J Mol Evol (1996) 42:476–480

JOURNAL OF MOLECULAR EVOLUTION

Letter to the Editor

Unbiased Estimation of Symmetrical Directional Mutation Pressure from Protein-Coding DNA

Lars S. Jermiin,[1,*] Peter G. Foster,[1] Dan Graur,[2] Roger M. Lowe,[3] Ross H. Crozier[3]

[1] Department of Biology, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada
[2] Department of Zoology, George S Wise Faculty of Life Science, Tel Aviv University, Ramat Aviv 69978, Israel
[3] School of Genetics and Human Variation, La Trobe University, Bundoora, Victoria 3083, Australia

Received: 5 May 1995 / Accepted: 30 November 1995

Abstract. The most generally applicable procedure for obtaining estimates of the symmetrical, or strand-nonspecific, directional mutation pressure (μ_D) on protein-coding DNA sequences is to determine the G+C content at synonymous codon sites (P_{syn}), and to divide P_{syn} by twice the arithmetic mean of the G+C content at synonymous codon sites of a large number of randomly generated, synonymously coding DNA sequences ($\bar{P}_{syn}$). Unfortunately, the original procedure yields biased estimates of P_{syn} and μ_D and is computationally expensive. We here present a fast procedure for estimating unbiased μ_D values. The procedure employs direct calculation of $\hat{P}_{syn}$ ($\approx\bar{P}_{syn}$) and two normalization procedures, one for $P_{syn} \leqslant \hat{P}_{syn}$ and another for $P_{syn} \geqslant \hat{P}_{syn}$. The normalization removes a bias sometimes caused by codons specifying arginine, asparagine, isoleucine, and leucine. Consequently, comparison of protein-coding genes that are translated using different genetic codes is facilitated.

Key words: Symmetrical directional mutation pressure — A+T pressure — G+C pressure — Synonymous codon sites — Nonsynonymous codon sites — Bias correction

* *Present address:* John Curtin School of Medical Research, Australian National University, Canberra, ACT 0200, Australia; e-mail: lars.jermiin@anu.edu.au
Correspondence to: L. S. Jermiin

Introduction

The molecular evolution of metazoan mitochondrial DNA has received much attention in recent years. Based on heterogenous patterns of base composition at various coding and noncoding parts of this DNA, it has been concluded that directional mutation pressure has had a significant impact on the evolution of metazoan mitochondrial DNA (Jukes and Bhushan 1986; Andersson and Kurland 1991; Asakawa et al. 1991, 1995; Osawa et al. 1992; Jermiin et al. 1994, 1995). Directional mutation pressure occurs when the rate of a nucleotide substitution (e.g., A → C) differs from that in the opposite direction (i.e., C → A). This definition is a generalization of the original definition by Sueoka (1962) which focuses on mutational bias in terms of A+T or G+C pressure.

Several methods for detecting directional mutation pressure have been proposed and include analysis of correlations between the G+C content at different codon sites (Jukes and Bhushan 1986) and between the G+C content of tRNAs, rRNAs, proteins, and spacers and that of the total genome (Muto and Osawa 1987). Other methods involve analysis of correlations between the relative abundance of particular amino acids and the G+C content of the complete genome (Sueoka 1961a,b), of silent sites (Collins and Jukes 1993), of different positions (D'Onofrio et al. 1991; Sueoka 1992), and of codon families (Crozier and Crozier 1993; Jermiin and

Crozier 1994). Recently, Jermiin et al. (1994) proposed the *synonymous sites approach,* according to which all positions in the DNA are divided into synonymous and nonsynonymous codon sites and used to calculate the G+C content at synonymous and nonsynonymous codon sites (P_{syn} and P_{non}, respectively). The G+C content at the synonymous codon sites is then used with a genetic code to calculate a normalized estimate of the symmetrical directional mutation pressure (μ_D). The advantage of this method is that the normalization, which removes a bias induced by threefold-degenerate codon families, facilitates comparing μ_D or P_{non} values from genes that are translated by different genetic codes. The method is computationally expensive. Nonetheless, it has been used successfully to detect and assess the effect of symmetrical directional mutation pressure on more than a hundred mitochondrial protein-coding genes (Jermiin et al. 1994).

Recently it became clear that the *synonymous sites approach* may lead to incorrect estimates of the G+C content at synonymous and nonsynonymous codon sites and hence of symmetrical directional mutation pressure. Analyzing a 1.2-kbp segment of the G+C-rich elongation factor 1α gene from *Giardia lamblia,* we obtained a μ_D value larger than 1.0, which is in disagreement with the theory of directional mutation pressure (Sueoka 1962). Moreover, we found that the sequence exclusively uses the CGN codon to specify arginine (with a frequency of 3.54%) whereas the A+T-rich *Entamoeba histolytica* equivalent exclusively uses the AGR codon (with a frequency of 4.06%).[1] This result is reason for concern because it implies that a synonymous substitution can occur at a site that is considered nonsynonymous according to the *synonymous sites approach,* and hence P_{non} values may be biased. The purpose of this paper is to identify and correct the problems with the *synonymous sites approach* so that it estimates correctly the P_{syn}, P_{non}, and μ_D values, and to improve the method so that μ_D can be calculated without the use of time-consuming computer simulations.

Unbiased Estimation of the μ_D Value

In order to identify the error in the *synonymous sites approach,* we generated four protein-coding DNA sequences (Table 1) and analyzed them using the DMP program (see Jermiin et al. 1994). When they were analyzed using the vertebrate mitochondrial genetic code (isoleucine is coded by a twofold-degenerate codon family), the μ_D values are identical to the P_{syn} values and fall between 0.0 and 1.0 (Table 2). However, when they were analyzed using the echinoderm mitochondrial genetic code (isoleucine is coded by a threefold-degenerate codon family), the μ_D values differ from the P_{syn} values in all but one case and may become as large as 1.5 (Table 2). The normalization thus works correctly in some cases because μ_D values for sequence 1 and 2 equal expected values (Jermiin et al. 1994). However, the μ_D value for sequence 4 is larger than 1.0. Thus, the result indicates the normalization proposed by Jermiin et al. (1994) works correctly under some conditions but fails under others.

Table 1. DNA sequences used in this study[a]

Sequence #	Sequence structure	P_{syn}
1	$(ATT\ ATT\ ATT\ ATT\ ATT\ ATT)_{100}$	0.0000
2	$(ATT\ ATT\ ATC\ ATT\ ATT\ ATC)_{100}$	0.3333
3	$(ATT\ ATC\ ATT\ ATC\ ATT\ ATC)_{100}$	0.5000
4	$(ATC\ ATC\ ATC\ ATC\ ATC\ ATC)_{100}$	1.0000

[a] The four DNA sequences, each containing 1,800 base pairs, were constructed so that they encode a sequence of isoleucine using any one of the genetic codes. The G+C content at the synonymous codon sites (P_{syn}) was calculated using the DMP program (Jermiin et al. 1994)

The normalization is intended to produce an unbiased μ_D value from a P_{syn} value which may be biased by threefold-degenerate codon families. We illustrate the normalization using a DNA sequence that encodes a polyisoleucine sequence. The synonymous sites of this sequence are allowed to mutate at random (thus μ_D = 0.5) and therefore will contain equal quantities of T and C if the sequence is translated by the vertebrate mitochondrial genetic code. The mean of the G+C content at synonymous codon sites of such randomly mutating, synonymously coding DNA sequences ($\bar{P}_{syn}$) is equal to 0.5; thus transformation of a P_{syn} value to its corresponding μ_D value is direct (Fig. 1A). If the same sequence is translated by the echinoderm mitochondrial genetic code, it will contain equal proportions of A, T, and C, and therefore $\bar{P}_{syn}$ will be equal to 0.333. Under such conditions ($\bar{P}_{syn}$ 0.5), direct transformation is inappropriate, and a transformation that takes into account the deviation of $\bar{P}_{syn}$ from 0.5 is required (note that $0.333 \leqslant \bar{P}_{syn} \leqslant 0.5$—this is a consequence of the structure of threefold-degenerate codon families). An expansion of the axis is needed for values of $P_{syn} \leqslant \bar{P}_{syn}$ whereas a compression of the axis is needed for values of $P_{syn} \geqslant \bar{P}_{syn}$ (Fig. 1B). With respect to the transformation by Jermiin et al. (1994), multiplying P_{syn} values by $1/(2\bar{P}_{syn})$ yields correct expansion of the axis for $P_{syn} \leqslant \bar{P}_{syn}$ (Fig. 1C). However, compression of the axis for $P_{syn} \geqslant \bar{P}_{syn}$ never occurs—the axis continues to expand and μ_D values larger than 1.0 can therefore be obtained (Fig. 1C).

Based on these results, we propose that the normalization uses the following two equations for generating unbiased estimates of μ_D (conditions in brackets):

[1] The elongation factor 1α genes in *Giardia lamblia* (Hashimoto et al. 1994) and *Entamoeba histolytica* (De Meester et al. 1991) were obtained from GenBank (accession numbers D14342, M92073, and M34256).

Table 2. Comparison of μ_D values obtained after normalization using the vertebrate and echinoderm mitochondrial genetic codes[a]

		Jermiin et al. (1994)		This study	
Sequence #	P_{syn}	μ_D (verteb.)	μ_D (echino.)	μ_D (verteb.)	μ_D (echino.)
1	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.3333	0.3333	0.5000	0.3333	0.5000
3	0.5000	0.5000	0.7500	0.5000	0.6250
4	1.0000	1.0000	1.5000	1.0000	1.0000

[a] The four sequences and the P_{syn} values are from Table 1. The μ_D values were obtained using the normalization proposed by Jermiin et al. (1994) and that proposed in this paper. For the present study, equation (1) was used for sequences 1 and 2 where equation (2) was used for sequences 3 and 4. Note that μ_D values for sequences 3 and 4 in the sixth column differ from corresponding values in the fourth column

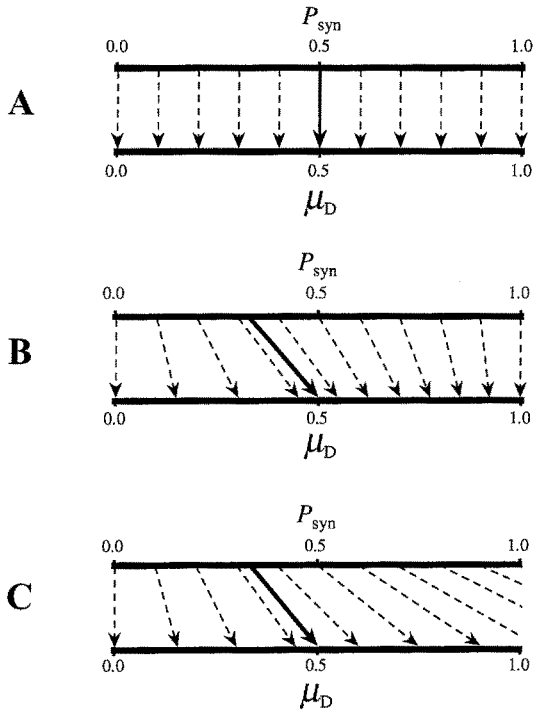

Fig. 1. Transformation of P_{syn} values to μ_D values—*arrows* illustrate relationships between P_{syn} values to their corresponding μ_D values (transformation of $\bar{P}_{syn}$ is represented by a *heavy arrow*). **A** Transformation is direct when $\bar{P}_{syn}$ 0.5, in which case μ_D P_{syn}. **B** Transformation is indirect when $\bar{P}_{syn}$ 0.5, in which case μ_D P_{syn}, except when P_{syn} equals 0.0 or 1.0. Indirect transformation requires expansion of the axis when $P_{syn} \leq \bar{P}_{syn}$ and compression of the axis when $P_{syn} \geq \bar{P}_{syn}$. **C** Indirect transformation proposed by Jermiin et al. (1994). Note that all values on the P_{syn} axis expand regardless of their position in relation to $\bar{P}_{syn}$, and therefore that μ_D values larger than 1.0 can be obtained.

$$\mu_D = \frac{P_{syn}}{2\bar{P}_{syn}} \qquad [\text{for all } P_{syn} \leq \bar{P}_{syn}] \quad (1)$$

$$(1-\mu_D) = \frac{(1-P_{syn})}{2(1-\bar{P}_{syn})} \qquad [\text{for all } P_{syn} \geq \bar{P}_{syn}] \quad (2)$$

Equations (1) and (2) have expanding and compressing properties, respectively, when $\bar{P}_{syn} < 0.5$, and therefore yield transformations between P_{syn} and μ_D values that are identical to those illustrated in Fig. 1B. Application of the revised normalization is illustrated in Table 2, and we conclude that normalization using equations (1) and (2) yields μ_D values that are in agreement with the theory of directional mutation pressure (Sueoka 1962).

Unbiased Estimation of the P_{non} Value

The *synonymous sites approach* is based on correct separation of sites in a protein-coding DNA sequence into synonymous and nonsynonymous codon sites. In order to achieve this partition the codon family was treated as a unit and sites within each codon family were defined either as synonymous or nonsynonymous (Jermiin et al. 1994). This means that only third codon sites are considered synonymous (except ATG and TGG in nuclear genes and mitochondrial genes of plants, AAG and ATG in mitochondrial genes of echinoderms, and ATG in mitochondrial genes of euascomycetes). The drawback of this approach is that first codon sites in codons specifying arginine or leucine sometimes are considered nonsynonymous when in fact they are synonymous. For example, two identically coding DNA sequences could have different P_{non} values if the first codon sites of their leucine codons contained T and C, respectively,[2] and this is considered undesirable.

We propose to include the first codon position of codons specifying arginine and leucine as a synonymous site whenever appropriate.[3] Our reason is that the genetic codes specify whether a single nucleotide substitution at those sites will allow a change between codon families to occur without also changing the corresponding protein sequence. The proposed change improves the *synonymous sites approach* because it eliminates a bias previously induced on P_{non} and because differences between

[2] Except yeast mitochondrial DNA, which only uses the TTR codon to specify leucine.

[3] This happens when codons specifying arginine are translated by the universal genetic code or the mitochondrial genetic codes in plants, yeast, or euascomycetes and when codons specifying leucine are translated by the universal genetic code or the mitochondrial genetic codes in vertebrates, arthropods, nematodes, echinoderms, plants, or euascomycetes.

P_{non} values now always will be associated with compositional differences among the corresponding protein sequences. We note that this classification of synonymous and nonsynonymous codon sites now is identical to the commonly used classification of degenerate and nondegenerate codon sites.

Direct Calculation of the μ_D Value

The determination of unbiased μ_D values relies on precise estimates of $\bar{P}_{syn}$. Previously, this has been achieved by time-consuming computer simulation. In order to improve precision and speed concurrently, we present a direct method for calculating $\bar{P}_{syn}$. Let C_2, C_3, C_4, and C_6 be clusters of 2, 3, 4, and 6 synonymous codons, respectively, and let N_2, N_3, N_4, and N_6 be the frequencies of C_2, C_3, C_4, and C_6, respectively.[4] The precise estimate of $\bar{P}_{syn}$—denoted $\hat{P}_{syn}$—can be calculated as

$$\hat{P}_{syn} = \frac{1}{2}\frac{N_4+N_2}{N_6+N_4+N_3+N_2} + \frac{1}{3}\frac{N_3}{N_6+N_4+N_3+N_2} + \frac{7}{12}\frac{N_6}{N_6+N_4+N_3+N_2} \quad (3)$$

where the factors one-half, one-third, and seven-twelfths denote the G+C content at randomly mutating (thus $\mu_D = 0.5$) synonymous codon sites in DNA sequences that contain exclusively C_2 and C_4, C_3, and C_6, respectively. Applying equation (3) to the sequences in Table 1, we get $\hat{P}_{syn} = 0.5$ if they are translated by the vertebrate mitochondrial genetic code and $\hat{P}_{syn} = 0.333$ if they are translated by the echinoderm mitochondrial genetic code. These two values are almost identical to their corresponding $\bar{P}_{syn}$ values, which were determined by simulation.

An Example

The unbiased μ_D value from a protein-coding DNA sequence can be calculated as follows. Determine the observed G+C content of the DNA sequence (P_{obs}) (exclude the stop codon), the G+C content of a synonymously coding sequence in which the synonymous sites are saturated with G and C (P_{max}), and the G+C content of a synonymously coding sequence in which the synonymous sites are saturated with A and T (P_{min}). Following Jermiin et al. (1994), the G+C content at the synonymous sites (P_{syn}) is then given as

$$P_{syn} = \frac{(P_{obs} - P_{min})}{(P_{max} - P_{min})} \quad (4)$$

The G+C content at synonymous sites of a randomly mutating, synonymously coding sequence ($\hat{P}_{syn}$) is then determined using equation (3), and $\hat{P}_{syn}$ is used in equations (1) and (2) (instead of $\bar{P}_{syn}$) to calculate the unbiased μ_D value. The variance of the μ_D value follows the binomial distribution and is given by

$$\sigma_b^2(\mu_D) = \frac{\mu_D(1-\mu_D)}{b} \quad (5)$$

where b is the number of synonymous codon sites in the sequence (Sueoka 1962).

We illustrate the calculation of μ_D and its variance by using the published part of the *Giardia lamblia* elongation factor 1α (Hashimoto et al. 1994). The values of P_{obs}, P_{min}, and P_{max} are equal to 0.6481, 0.3316, and 0.6498, respectively,[5] and hence P_{syn} equals 0.9947. The values of N_2, N_3, N_4, and N_6 are equal to 161, 30, 150, and 37, respectively, and therefore $\hat{P}_{syn}$ equals 0.4949. Using the normalization of Jermiin et al. (1994), we get $\mu_D = 1.0049$ and $\sigma_b^2(\mu_D) = -1.3 \times 10^{-5}$, which is impossible according to statistical theory and in conflict with the theory of directional mutation pressure (Sueoka 1962). However, normalization as proposed in this paper yields $\mu_D = 0.9946$ and $\sigma_b^2(\mu_D) = 1.4 \times 10^{-5}$.

Acknowledgments. We thank D. Hickey, S. Wang, and an anonymous referee for constructive comments on the manuscript. L.S.J. and P.G.F. were supported by an NSERC Canada Research Grant to D. Hickey, and R.H.C.'s research was supported by the Ian Potter Foundation and the Australian Research Council's Special Investigator Award.

References

Andersson SGE, Kurland CG (1991) An extreme codon preference strategy: codon reassignment. Mol Biol Evol 8:530–544

Asakawa S, Kumazawa Y, Araki T, Himeno H, Miura K-I, Watanabe K (1991) Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. J Mol Evol 32:511–520

Asakawa S, Himeno S, Miura K-I, Watanabe K (1995) Nucleotide sequence and gene organization of the starfish *Asterina pectinifera* mitochondrial genome. Genetics 140:1047–1060

Collins DW, Jukes TH (1993) Relationship between G+C in silent sites of codons and amino acid composition of human proteins. J Mol Evol 36:201–213

Crozier RH, Crozier YC (1993) The mitochondrial genome of the honeybee *Apis mellifera:* complete sequence and genome organization. Genetics 113:97–117

De Meester F, Bracha R, Huber M, Keren Z, Rozenblatt S, Mirelman D (1991) Cloning and characterization of an unusual elongation

[4] The two codon families specifying serine cannot be linked by a single nucleotide substitution and their codon usage frequencies are therefore included in N_2 and N_4 (depending on the genetic code). The same applies to the codon usage frequency of codons specifying threonine in yeast mitochondrial DNA.

[5] These values were obtained using the DMP program. An IBM-compatible version of DMP is available from L.S.J.

factor-1 alpha cDNA from *Entamoeba histolytica.* Mol Biochem Parasitol 44:23–32

D'Onofrio GD, Mouchiroud D, Aïssani B, Gautier C, Bernardi G (1991) Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J Mol Evol 32:504–510

Hashimoto T, Nakamura Y, Nakamura F, Shirakura T, Adachi J, Goto N, Okamoto K-I, Hasagawa M (1994) Protein phylogeny gives a robust estimation for early divergences of Eukaryotes: phylogenetic place of a mitochondrial-laking protozoan, *Giardia lamblia.* Mol Biol Evol 11:65–71

Jermiin LS, Crozier RH (1994) The cytochrome *b* region in the mitochondrial DNA of the ant *Tetraponera rufoniger:* sequence divergence in hymenoptera may be associated with nucleotide content. J Mol Evol 38:282–294

Jermiin LS, Graur D, Lowe RM, Crozier RH (1994) Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome *b* genes. J Mol Evol 39:160–173

Jermiin LS, Graur D, Crozier RH (1995) Evidence from analyses of intergenic regions for strand-specific directional mutation pressure in metazoan mitochondrial DNA. Mol Biol Evol 12:558–563

Jukes TH, Bhushan V (1986) Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. J Mol Evol 24:39–44

Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA 84:166–169

Osawa S, Jukes TH, Watanabe K, Muto A (1992) Recent evidence for evolution of the genetic code. Microbiol Rev 56:229–264

Sueoka N (1961a) Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. Proc Natl Acad Sci USA 47:1141–1149

Sueoka N (1961b) Compositional correlation between deoxyribonucleic acid and protein. Cold Spring Harb Symp Quant Biol 26:35–43

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci USA 48:582–592

Sueoka N (1992) Directional mutation pressure, selective constraints, and genetic equilibria. J Mol Evol 34:95–114

J Mol Evol (1997) 44:468

JOURNAL OF MOLECULAR EVOLUTION

Erratum

Unbiased Estimation of Symmetrical Directional Mutation Pressure from Protein-Coding DNA

Lars S. Jermiin,[1,*] Peter G. Foster,[1] Dan Graur,[2] Roger M. Lowe,[3] Ross H. Crozier[3]

[1] Department of Biology, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada
[2] Department of Zoology, George S. Wise Faculty of Life Science, Tel Aviv University, Ramat Aviv 69978, Israel
[3] School of Genetics and Human Variation, La Trobe University, Bundoora, Victoria 3083, Australia

Re: J Mol Evol (1996) 42:476–480. Please note the following corrections to this article:

Page 477, 2nd column: ''Under such conditions ($\overline{P}_{syn}$ 0.5), direct transformation . . . '' should have been ''Under such conditions ($\overline{P}_{syn} \neq 0.5$), direct transformation . . . ''.

Page 478, Fig. 1: ''Transformation is direct when $\overline{P}_{syn}$ 0.5, in which case μ_D P_{syn}.'' should have been ''Transformation is direct when $\overline{P}_{syn} = 0.5$, in which case $\mu_D = P_{syn}$.''

Page 478, Fig. 1: ''Transformation is indirect when $\overline{P}_{syn}$ 0.5, in which case μ_D P_{syn}, except when . . .'' should have been ''Transformation is indirect when $\overline{P}_{syn} \neq 0.5$, in which case $\mu_D \neq P_{syn}$, except when . . . ''.

* *Present address:* John Curtin School of Medical Research, Australian National University, Canberra, ACT 0200, Australia; e-mail: lars.jermiin@anu.edu.au
Correspondence to: L.S. Jermiin