Uncorrected Nucleotide Bias in mtDNA Can Mimic the Effects of Positive Darwinian Selection

Mihai Albu, Xiang Jia Min,†‡ Donal Hickey,‡ and Brian Golding**

*Department of Biology, McMaster University, Hamilton, ON L8S 4K1, Canada; †Department of Biological Sciences, Youngstown State University, Youngstown; and ‡Department of Biology, Concordia University, Montréal, QC H3G 1M8, Canada

The relative rates of nucleotide substitution at synonymous and nonsynonymous sites within protein-coding regions have been widely used to infer the action of natural selection from comparative sequence data. It is known, however, that mutational and repair biases can affect rates of evolution at both synonymous and nonsynonymous sites. More importantly, it is also known that synonymous sites are particularly prone to the effects of nucleotide bias. This means that nucleotide biases may affect the calculated ratio of substitution rates at synonymous and nonsynonymous sites. Using a large data set of animal mitochondrial sequences, we demonstrate that this is, in fact, the case. Highly biased nucleotide sequences are characterized by significantly elevated *d*N/*d*S ratios, but only when the nucleotide frequencies are not taken into account. When the analysis is repeated taking the nucleotide frequencies at each codon position into account, such elevated ratios disappear. These results suggest that the recently reported differences in *d*N/*d*S ratios between vertebrate and invertebrate mitochondrial sequences could be explained by variations in mitochondrial nucleotide frequencies rather than the effects of positive Darwinian selection.

Introduction

Protein-coding sequences are subject to various forms of natural selection, and the effects of such selection can be inferred from comparative sequence analysis. The most obvious pattern that emerges from a comparison of orthologous protein-coding sequences is that the third position of codons usually shows the highest levels of nucleotide variability, whereas the second codon position is the least variable. This pattern can be explained by the fact that nucleotide changes at the third codon position are often synonymous—encoding the same amino acid—whereas second position changes are always nonsynonymous (Kimura 1977). Consequently, the second codon position is expected to be under greater selective constraint than the third position.

A more detailed study of the patterns of nucleotide substitution at synonymous and nonsynonymous sites is often used to infer different forms of natural selection. Specifically, the ratio of these rates can be used to distinguish between purifying selection and positive Darwinian selection (Hughes and Nei 1988; Li 1993; Yang 1998; Yang and Nielsen 2002; Nei 2005). Although the inferred rate of nucleotide substitution at synonymous sites is usually higher than that at nonsynonymous sites, the ratio of these rates also varies between genes and between organisms. These variations can, in turn, be used to infer variations in the intensity and direction of natural selection acting at the nonsynonymous sites. These inferences are based on the assumption that substitutions at synonymous sites are largely selectively neutral or at least under small selective constraint relative to the intensity of selection at nonsynonymous sites. Given this assumption, a very low proportion of nonsynonymous mutations can be interpreted as a reflection of intense purifying selection maintaining a functional amino acid sequence (for a review, see Meiklejohn et al. 2007).

Key words: *d*N/*d*S ratios, nucleotide bias, synonymous sites, nonsynonymous sites, PAML, CODEML, mtDNA.

E-mail: golding@mcmaster.ca.

Mol. Biol. Evol. 25(12):2521–2524. 2008
doi:10.1093/molbev/msn224
Advance Access publication October 8, 2008

Much attention has been focused on a situation that, a priori, seems unlikely, that is, those cases where the substitution rate at nonsynonymous sites is significantly greater than at the weakly selected synonymous sites (e.g., Hill and Hastie 1987; Hughes and Nei 1988; Yang and Nielsen 2002; Swanson et al. 2003). The reason for this attention is that a high proportion of nonsynonymous substitutions implies the action of diversifying selection between lineages, also known as positive Darwinian selection.

Although the logic behind these inferences is straightforward, the accurate calculation of these ratios of substitution rates can be confounded by a number of factors in practice. For example, there is an obvious problem of site saturation when calculating substitution rates from observed sequence differences (Gojobori 1983). In addition and of more relevance here is the fact that nucleotide frequencies may vary between lineages and that such nucleotide biases may affect the calculated rate of change at synonymous sites (Aris-Brosou and Bielawski 2006; Friedman and Hughes 2007).

In this study, we chose a set of well-studied, single-copy orthologous genes encoded by animal mitochondrial genomes. Not only are these genes well characterized with regard to their function but their coding sequences are also known to show a variety of significant nucleotide and amino acid biases (Jermiin and Crozier 1994; Perna and Kocher 1995; Foster et al. 1997; Rand and Kann 1998; Reyes et al. 1998; Gibson et al. 2005). Thus, these sequences provide a well-defined data set for the study of nucleotide biases on the inference of nucleotide substitution rates at synonymous and nonsynonymous sites.

Methods

Following the methodologies of Bazin et al. (2006), we selected from the NCBI Ref Seq organelle genome database (v. 160.0, release June 2007) the completed mitochondrial DNA (mtDNA) genomes of species that belong to the following class/phylum classifications: Amphibia, Chondrichthyes, Sauropsida, Mammalia, Teleostei, Insecta, Crustacea, Mollusca, Nematoda, and Chelicerata (supplementary table ST1, Supplementary Material online).

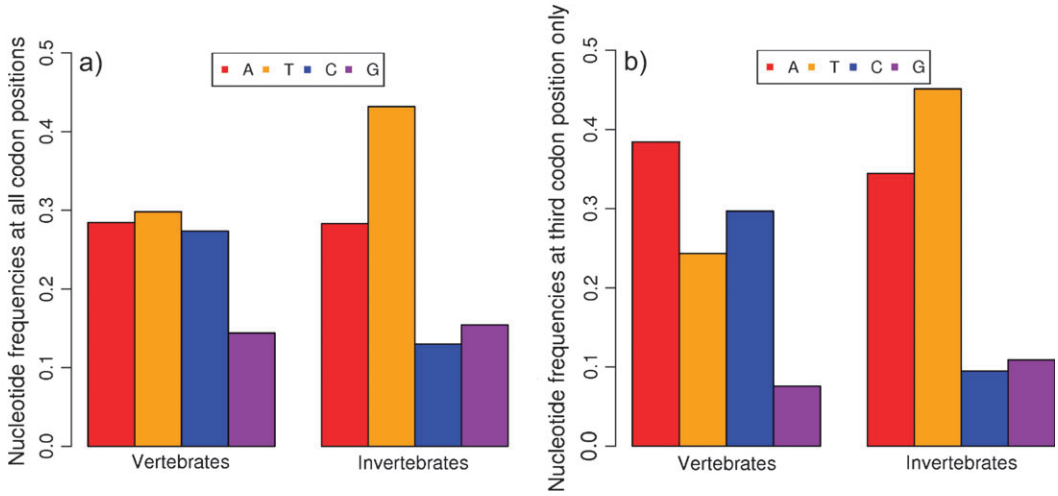

FIG. 1.—Differences in average nucleotide content between vertebrate and invertebrate mitochondrial coding sequences. (*a*) Shows the average frequency of each of the four nucleotides (A = red, T = orange, C = blue, and G = violet) for all three codon positions. (*b*) Shows the same data for the third codon position only. The corresponding data for subgroups within the vertebrates and invertebrates are shown in supplementary figure S1 (Supplementary Material online).

For each mtDNA genome, all 13 protein-coding gene sequences were extracted. For each taxonomic group, a single *d*N/*d*S ratio was calculated from the average for all genes in that group. We restricted the analyses to the family classifications that have records of at least two species. For each family and each gene, coding sequences and protein sequences were extracted from the mtDNA genome and aligned using ClustalW (Thompson et al. 1997). The aligned sequences were used in the CODEML program from the PAML package (Yang 2007) to estimate *d*N/*d*S. Analyses were performed with the following parameters: estimations using pairwise comparisons (runmode = −2) and estimations using comparisons that take into account the phylogenetic history (runmode = 0); omega (measuring the *d*N/*d*S ratio) and kappa (measuring the transition to transversion ratio) were estimated from the sequence data based on two alternate models: 1) codon frequencies were estimated based on the assumption of equal nucleotide frequencies at all codon positions (model CF = 0) or 2) codon frequencies estimated based on the actual nucleotide frequencies at each of the three codon positions (model CF = 2). The mitochondrial genetic code was adjusted for each of the species according to the listings in GenBank. Manual inspection of data sets removed pairs of sequences with no synonymous divergence ($dS = 0$) and pairs of sequences with very large divergence ($t > 10$, where t is the expected number of nucleotide substitutions per codon).

The *d*N/*d*S values for each group are reported as averages over all the constituent families and genes. The analysis of the "effective number of codons" (ENC; Wright 1990) was performed using CodonW (http://codonw.sourceforge.net). All plots and statistical analyzes were created using R (http://www.r-project.org).

Nucleotide divergences at synonymous and nonsynonymous sites were calculated taking the phylogenetic history of the sequences into account. For each family and each gene, a phylogenetic history was constructed using the Neighbor-Joining algorithm from the PHYLIP package (Felsenstein 1989) with distances calculated according to Hasegawa–Kishino–Yano model of substitutions implemented in the Tree-Puzzle package (Schmidt et al. 2002). The inferred phylogenies were used as input to the CODEML program (Yang 2007). For comparison, the analysis was repeated using a nonphylogenetic, pairwise approach and the results from the two methods were compared.

Results

First, we compared the average nucleotide content of mitochondrial coding sequences between vertebrates and invertebrates (see fig. 1). In this figure, we show the frequencies of the four nucleotides separately because this captures both the AT/GC bias as well as the strand biases (AT and GC skews). Our results show that although mitochondrial sequences are, in general, GC poor, they also tend to have GC skews (unequal frequencies of C and G) and AT skews (unequal frequencies of A and T). By comparing figure 1*a* and *b*, we can see that, as expected, the nucleotide biases are much more obvious at the largely synonymous third codon positions. From figure 1*b*, we can also see that there are marked differences in nucleotide frequencies between the vertebrates and invertebrates. Among the vertebrates, the GC skews are generally negative (C more frequent than G on the coding strand), whereas the AT skews are positive (A more frequent than T); among the invertebrates, however, the GC skews are not markedly negative but, in contrast to the vertebrates, the AT skews are negative (see fig. 1*b*). The most striking difference is that the vertebrates have a significantly lower frequency of T nucleotides than invertebrates ($P < 0.001$), although they have a significantly higher frequency of C nucleotides ($P < 0.0001$). The variation in nucleotide content within both vertebrates and invertebrates is indicated in supplementary figure SF1 (Supplementary Material online).

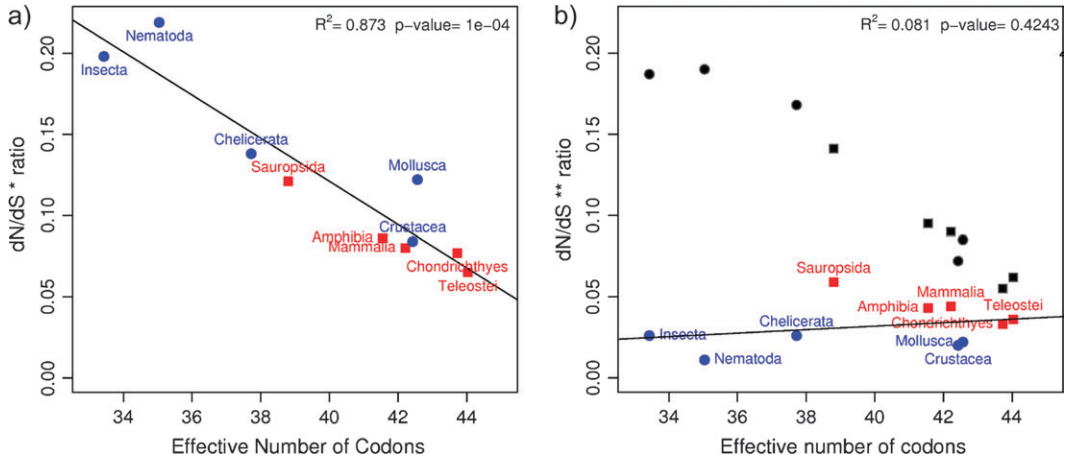

FIG. 2.—Correlation between the *d*N/*d*S estimations and the ENC. (*a*) Shows the correlation between the published *d*N/*d*S ratios (Bazin et al. 2006) and the ENC. Pearson's squared correlation coefficient $R^2 = 0.873$. (*b*) Shows the correlation between the *d*N/*d*S ratios and the ENC, when the nucleotide frequencies at the three codon positions are taken into account. In this case, there is no obvious correlation between the *d*N/*d*S ratios and the ENC ($R^2 = 0.081$). For comparison, we also show the results assuming equal nucleotide frequencies (black symbols). Blue circles indicate the values for the invertebrate groups, and red squares indicate the values for the vertebrate groups.

Next we asked if these differences in nucleotide content might affect the calculation of the ratio of nucleotide divergences at synonymous and nonsynonymous sites, that is, *d*N/*d*S ratios. To answer this question, we first plotted the published *d*N/*d*S ratios (Bazin et al. 2006) against the degree of nucleotide bias for the corresponding groups of vertebrates and invertebrates. As an index of nucleotide bias, we used the ENC as this captures both the GC bias and the strand-specific skews (Wright 1990). The results show that there is indeed a highly significant negative correlation (correlation coefficient $R^2 = 0.87$; $P = 1 \times 10^{-04}$) between the degree of nucleotide bias and the *d*N/*d*S ratio (see fig. 2*a*). A similar result can be obtained if GC content is used as the index of nucleotide bias (see supplementary fig. SF2a, Supplementary Material online). Based on these results, it would appear that the reported higher *d*N/*d*S ratios among the invertebrates might be explained by the fact that they have a greater average degree of nucleotide bias in their mtDNA.

These results provide circumstantial evidence that nucleotide biases can affect the calculation of *d*N/*d*S ratios. To test if these correlations are indeed due to the variations in nucleotide content, we recalculated the *d*N/*d*S ratios using a variety of different approaches. Specifically, we were interested in finding out if some methods were more sensitive to the effects of nucleotide bias than others. First, we compared the results of tree-based and pairwise methods for calculating the nucleotide divergences at synonymous and nonsynonymous sites (see Methods). The results are shown in supplementary figure SF3 (Supplementary Material online). There is a very good agreement between the two methods when the sequences are only moderately biased (overall correlation coefficient $R^2 = 0.892$; $P < 0.0001$). For the extremely biased groups such as the Nematoda and the Insecta, however, the tree-based method is less affected by the bias. Nevertheless, it should be noted that the improvement over the simple pairwise method is only moderate.

Our second approach was to repeat the analysis, while explicitly taking the nucleotide content at each codon position into account when calculating the values of *d*N and *d*S. This option is available within the PAML package (Yang 2007). The results are shown in figure 2*b*. In this case, we see that the negative correlation between the *d*N/*d*S ratio and the degree of nucleotide bias has completely disappeared. Moreover, the values for the invertebrates are now, if anything, slightly less than those for the vertebrates. This result shows that the reported average difference in *d*N/*d*S ratios between vertebrates and invertebrates is simply due to an average difference in the degree of nucleotide bias between the two groups (see also supplementary fig. SF2b, Supplementary Material online). This confirms that nucleotide bias can yield artificially high *d*N/*d*S ratios for biased sequences and that the degree of inflation is directly related to the severity of the nucleotide bias (see fig. 3).

Conclusion

The ratio of nonsynonymous to synonymous substitutions (*d*N/*d*S) is widely used to infer the action of natural selection from comparative sequence data. The results presented here demonstrate that these ratios can be affected significantly by nucleotide bias. This means that extreme caution should be used when comparing results between taxa that differ in their nucleotide contents. Specifically, those lineages that have more biased sequences (measured as GC bias and/or strand asymmetry) yield higher *d*N/*d*S ratios. This is caused largely by an apparent reduction in *d*S values among biased sequences (Blouin et al. 1998; Blouin 2000; Aris-Brosou and Bielawski 2006).

In summary, we have shown that strong biases in nucleotide content among mitochondrial sequences can yield artificially elevated *d*N/*d*S ratios. Consequently, the nucleotide frequencies at each codon position must be taken into

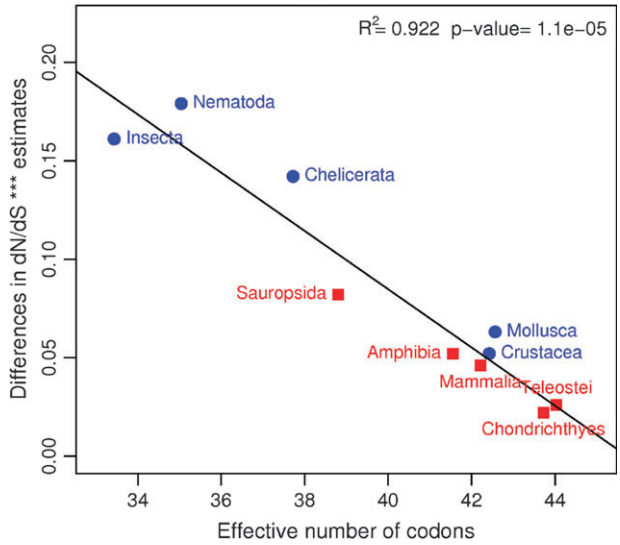

FIG. 3.—The effect of model choice on the magnitude of dN/dS estimates. The difference between the results using either 1) a model that assumes equal nucleotide frequencies in all groups or 2) a model that estimates codon frequencies from the average nucleotide frequencies at the three codon positions. Blue circles indicate the values for the invertebrate groups, and red squares indicate the values for the vertebrate groups. There is a strong negative correlation between the magnitude of the differences and the ENC (Pearson's squared correlation coefficient $R^2 = 0.922$).

account in order to obtain realistic estimates of dN/dS ratios.

Supplementary Material

Supplementary figures S1 and SF1–SF3 and table ST1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors express their thanks to the reviewers for their helpful comments in improving the presentation of this work. This work was supported through funding to the Canadian Barcode of Life Network from Genome Canada through the Ontario Genomics Institute, Natural Science and Engineering Research Council of Canada, and other sponsors listed at http://www.BOLNET.ca.

Literature Cited

Aris-Brosou S, Bielawski JP. 2006. Large-scale analyses of synonymous substitution rates can be sensitive to assumptions about the process of mutation. Gene. 378:58–64.

Bazin E, Glémin S, Galtier N. 2006. Population size does not influence mitochondrial genetic diversity in animals. Science. 312:570–572.

Blouin MS. 2000. Neutrality tests on mtDNA: unusual results from nematodes. J Hered. 91:156–158.

Blouin MS, Yowell CA, Courtney CH, Dame JB. 1998. Substitution bias, rapid saturation, and the use of mtDNA for nematode systematics. Mol Biol Evol. 15:1719–1727.

Felsenstein J. 1989. PHYLIP—phylogeny inference package (version 3.2). Cladistics. 5:164–166.

Foster PG, Jermiin LS, Hickey DA. 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J Mol Evol. 44:282–288.

Friedman R, Hughes AL. 2007. Likelihood-ratio tests for positive selection of human and mouse duplicate genes reveal non-conservative and anomalous properties of widely used methods. Mol Phylogenet Evol. 42:388–393.

Gibson A, Gowri-Shankar V, Higgs PG, Rattray M. 2005. A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. Mol Biol Evol. 22:251–264.

Gojobori T. 1983. Codon substitution in evolution and the "saturation" of synonymous changes. Genetics. 105:1011–1027.

Hill RE, Hastie ND. 1987. Accelerated evolution in the reactive centre regions of serine protease inhibitors. Nature. 326:96–99.

Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature. 335:167–170.

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.

Kimura M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature. 267:275–276.

Li W-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol. 36:96–99.

Meiklejohn CD, Montooth KL, Rand DM. 2007. Positive and negative selection on the mitochondrial genome. Trends Genet. 23:259–263.

Nei M. 2005. Selectionism and neutralism in molecular evolution. Mol Biol Evol. 22:2318–2342.

Perna NT, Kocher TD. 1995. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J Mol Evol. 41:353–358.

Rand DM, Kann LM. 1998. Mutation and selection at silent and replacement sites in the evolution of animal mitochondrial DNA. Genetica. 102–103:393–407.

Reyes A, Gissi C, Pesole G, Saccone C. 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol Biol Evol. 15:957–966.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18: 502–504.

Swanson WJ, Nielsen R, Yang Q. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol. 20:18–20.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25:4876–4882.

Wright F. 1990. The 'effective number of codons' used in a gene. Gene. 87:23–29.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 15:568–573.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 19:908–917.

Dan Graur, Associate Editor

Accepted September 30, 2008