

DNA Asymmetric Strand Bias Affects the Amino Acid Composition of Mitochondrial Proteins

Xiang JIA MIN and Donal A. HICKEY*

Department of Biology, Concordia University, 7141 Sherbrooke West, Montreal, Quebec, Canada H4B 1R6

(Received 20 July 2007; accepted 4 October 2007; published online 1 November 2007)

Abstract

Variations in GC content between genomes have been extensively documented. Genomes with comparable GC contents can, however, still differ in the apportionment of the G and C nucleotides between the two DNA strands. This asymmetric strand bias is known as GC skew. Here, we have investigated the impact of differences in nucleotide skew on the amino acid composition of the encoded proteins. We compared orthologous genes between animal mitochondrial genomes that show large differences in GC and AT skews. Specifically, we compared the mitochondrial genomes of mammals, which are characterized by a negative GC skew and a positive AT skew, to those of flatworms, which show the opposite skews for both GC and AT base pairs. We found that the mammalian proteins are highly enriched in amino acids encoded by CA-rich codons (as predicted by their negative GC and positive AT skews), whereas their flatworm orthologs were enriched in amino acids encoded by GT-rich codons (also as predicted from their skews). We found that these differences in mitochondrial strand asymmetry (measured as GC and AT skews) can have very large, predictable effects on the composition of the encoded proteins.

Key words: mitochondrion; strand asymmetry; amino acids; codon usage

1. Introduction

Genome-wide biases in nucleotide content have been extensively studied in a wide variety of organisms.¹ During the past decade, there has been accumulating evidence that these variations at the DNA level can result in parallel changes in the frequencies of amino acids in the encoded proteins. For example, the GC content of bacterial genomes has been shown to influence the amino acid composition of the proteome.^{2,3}

In addition to the variations in GC content, however, bacterial genomes can also show significant compositional asymmetry between the two DNA strands.^{4,5} This strand asymmetry is usually measured as GC or AT skew⁶ (see Methods) and it can be due to such factors

as different substitutional patterns between the leading and lagging strands during replication.^{4,7,8} By comparing the amino acid compositions of proteins encoded on the leading strand with those encoded on the lagging strand, it has been shown that these nucleotide skews can also affect the amino acid composition of bacterial proteins.^{9,10}

Among eukaryotes, the correlations between nucleotide content and amino acid compositions have been studied primarily in animal mitochondrial genomes. Animal mitochondria show considerable variation between species in both GC content and GC skew.⁶ It has been shown that the variations in GC content affect the amino acid composition of the encoded proteins¹¹ but the effects of variations in nucleotide skew between genomes have not been studied. The recent availability of a very large number of completely sequenced mitochondrial genomes allows us to fill this gap. Mammalian mitochondria, including those of humans, are characterized by negative GC skews and

Edited by Dr. Katsumi Isono

* To whom correspondence should be addressed. Tel. +1 514-848-2424 ext 5184. Fax. +1 514-848-2881. E-mail: dhickey@alcor.concordia.ca

positive AT skews, i.e. the major coding strand of mammalian mitochondria is relatively rich in the nucleotides C and A, and correspondingly poor in G and T. For instance, as noted by Perna and Kocher (1995),⁶ although the human mitochondrial genome contains more than 40% GC pairs, the frequency of G on the coding strand is only 5%. The strength of the skew varies between species, and some species of invertebrate show opposite skews to those found in mammals, i.e. the coding strand is rich in G and poor in C.⁶

In recent years, there has been an explosion of data on mitochondrial whole genome sequences. Since our goal was to measure the effect of strand asymmetry on amino acid composition, we chose a group of species, the Platyhelminthes (flatworms), in which the GC and AT skews are opposite to those seen in mammals.^{12,13} Among the flatworms, the coding sequences are rich in G and T and correspondingly poor in C and A, i.e. a complete contrast to the patterns seen in mammals. For example, the main mitochondrial coding strand in *Schistosoma mansoni* (a flatworm) contains only 6.7% C, but it contains 25% G.¹² Despite the contrasting GC and AT skew patterns, however, both mammals and flatworms contain essentially the same set of homologous mitochondrial genes and they span similar ranges of GC content. Thus a comparison between these two groups provides us with an opportunity to assess the potential effects of DNA strand bias on the amino acid composition of a well-characterized set of orthologous proteins.

2. Methods

We downloaded all of the publicly available complete mitochondrial (mt) genome (mtDNA) sequences from both mammals and Platyhelminthes from the NCBI RefSeq organelle genome database (http://www.ncbi.nlm.nih.gov/genomes/ORGANELLES/mztax_short.html) (released in July, 2006). There were a total of 170 mt genomes from mammals and 13 mt genomes from Platyhelminthes (Supplementary Table S1).

The nucleotide frequencies and compositional biases of mtDNA sequences were calculated from the '+'-strand (the major coding strand used by NCBI for annotation) for each species. The GC and AT asymmetry is measured in terms of GC- and AT-skews according to the following formulae given in Perna and Kocher⁶: GC-skew = $(G - C)/(G + C)$; AT-skew = $(A - T)/(A + T)$, where C, G, A, and T are the occurrences of the four nucleotides.

We predicted the amino acid compositions based on the mitochondrial genetic code (Table 1) by partitioning the mitochondrial codons into CA-rich codons, GT-rich codons, and other codons. The frequency of synonymous codon usage was measured by the nucleotide content of G + T or C + A at the third codon positions of fourfold degenerate codon families: GGN (Glycine), GTN (Valine), CGN (Arginine), ACN (Threonine), GCN (Alanine), and CCN (Proline) (Supplementary Fig. S1).

The statistical significance of the average differences in amino acid composition between the two groups of species was scored using a Student's *t*-test.

Table 1. Partition of the vertebrate mitochondrial genetic code into GT-rich, CA-rich and 'other' codons

| | | | |
|----------------------------|---------------------------|---------------------------|---------------------------|
| TTT F Phe (22, 103) | TCT S Ser (20, 28) | TAT Y Tyr (28, 53) | TGT C Cys (4, 32) |
| TTC F Phe (27, 10) | TCC S Ser (20, 3) | TAC Y Tyr (23, 8) | TGC C Cys (6, 4) |
| TTA L Leu (27, 66) | TCA S Ser (26, 11) | TAA * Ter (20, 4) | TGA W Trp (16, 15) |
| TTG L Leu (6, 56) | TCG S Ser (5, 6) | TAG * Ter (12, 5) | TGG W Trp (5, 16) |
| CTT L Leu (21, 12) | CCT P Pro (28, 12) | CAT H His (24, 14) | CGT R Arg (4, 10) |
| CTC L Leu (21, 2) | CCC P Pro (22, 1) | CAC H His (23, 2) | CGC R Arg (6, 1) |
| CTA L Leu (42, 4) | CCA P Pro (23, 5) | CAA Q Gln (25, 3) | CGA R Arg (8, 2) |
| CTG L Leu (9, 6) | CCG P Pro (5, 3) | CAG Q Gln (8, 4) | CGG R Arg (3, 3) |
| ATT I Ile (35, 48) | ACT T Thr (26, 16) | AAT N Asn (27, 21) | AGT S Ser (7, 29) |
| ATC I Ile (29, 4) | ACC T Thr (23, 2) | AAC N Asn (25, 2) | AGC S Ser (15, 3) |
| ATA M Met (33, 33) | ACA T Thr (30, 4) | AAA K Lys (24, 13) | AGA * Ter (6, 13) |
| ATG M Met (12, 30) | ACG T Thr (6, 4) | AAG K Lys (6, 16) | AGG * Ter (7, 12) |
| GTT V Val (7, 58) | GCT A Ala (12, 17) | GAT D Asp (9, 22) | GGT G Gly (6, 39) |
| GTC V Val (7, 5) | GCC A Ala (17, 2) | GAC D Asp (10, 1) | GGC G Gly (11, 3) |
| GTA V Val (14, 22) | GCA A Ala (14, 3) | GAA E Glu (15, 8) | GGA G Gly (14, 10) |
| GTG V Val (3, 23) | GCG A Ala (1, 4) | GAG E Glu (5, 12) | GGG G Gly (4, 15) |

GT-rich codons (italic) include GT, TG, GG, TT codons at the first two codon positions. TTA and TTG are excluded as there are other codons for Leu. CA-rich codons (bold) include CA, AC, CC, AA codons at the first two codon positions. Different codon assignments in Mammals and Platyhelminthes are underlined. In Platyhelminthes, AGA and AGG code for Ser, ATA for Ile, and AAA for Asn. The numbers following each codon are codon usage per thousand codons in the 11 conserved proteins of mammals (the first number) and Platyhelminthes (the second number).

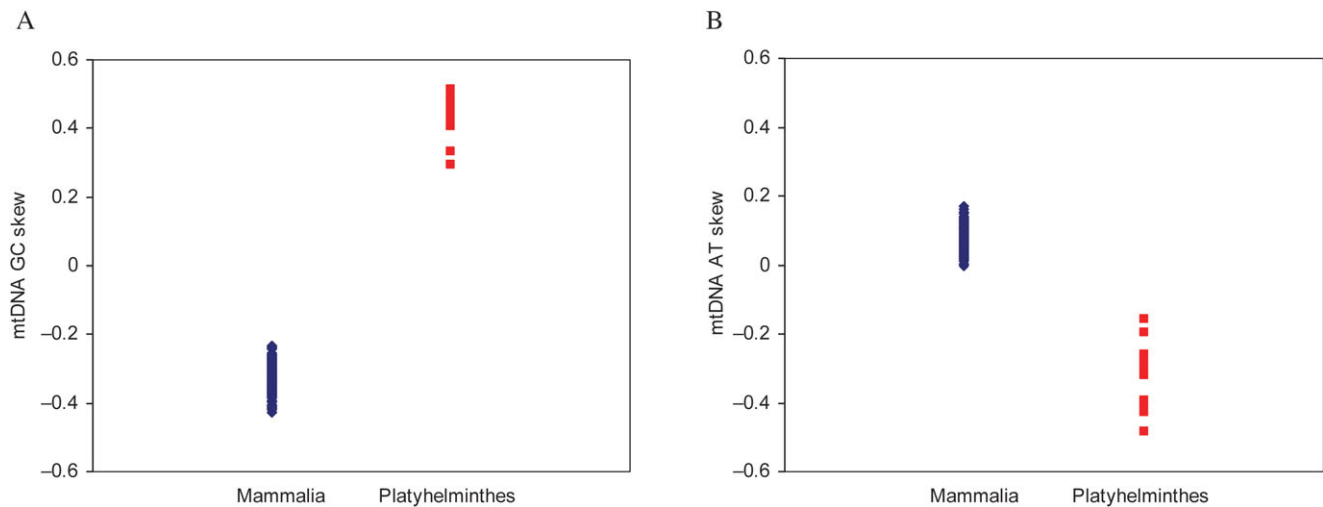


Figure 1. Contrasting patterns of DNA strand asymmetry in the mitochondrial coding sequences of mammals and Platyhelminthes. (A) The GC skew, $(G - C)/(G + C)$, is negative for all mammalian species (shown in blue), and positive for all Platyhelminthes species (shown in red). (B) The AT skew, $(A - T)/(A + T)$, is positive for mammals (shown in blue) and negative for Platyhelminthes (shown in red).

3. Results and discussion

We first confirmed the published reports (see Introduction) of contrasting GC and AT skews in the mitochondria of mammals and flatworms. As can be seen from Fig. 1, there is a negative GC skew, and a positive AT skew, in the major coding strand among the mammalian species, whereas the opposite is true among the flatworms. Despite the variations between species within both groups, there is a large average difference between them and this difference between the groups is statistically highly significant ($P < 0.0001$). As previously noted by Perna and Kocher⁶ and Le et al.,¹² the major coding sequence of mammalian mtDNA is relatively rich in C and A, whereas G and T are much more common in the flatworm coding sequences.

Despite their differences in strand asymmetry, the major coding strands of the two groups encode essentially the same set of 11 conserved mitochondrial proteins including cytochrome b, three subunits of cytochrome c oxidase (subunit 1, 2, and 3), six subunits of NADH dehydrogenase (subunit 1, 2, 3, 4, 4L, and 5), and ATP synthase F0 subunit 6 (ATP6). The exception is that NADH dehydrogenase subunit 6 (ND6) is encoded on the major coding strand in flatworms, but is encoded on the opposite strand in mammals.

Given the contrasting patterns of mitochondrial DNA strand asymmetry between these two groups of animals, we wished to investigate if there was a corresponding difference between the two groups in the frequencies of encoded amino acids. Specifically, because of their negative GC skews and positive AT skews (which reduce the frequencies of G and T nucleotides on the coding strand), we expected the mammalian coding strands to encode proteins that are relatively low in the proportions of cysteine (C), Valine (V), phenylalanine (F), Glycine

(G), and Tryptophan (W), all of which are encoded by GT-rich codons (Table 1). This is exactly what we observed (Fig. 2A). On average, the combined proportions of these five amino acids among the mammals are approximately half the value observed in the flatworm orthologs. Since these GC and AT skews result in a corresponding enrichment of C and A nucleotides on the mammalian coding strand, we expected the mammalian proteins to show a corresponding relative increase in the proportions of Glutamine (G), Threonine (T), Proline (P), Histidine (H), Asparagine (N), and Lysine (K), all of which are encoded by CA-rich codons. Again, this prediction is borne out (Fig. 2B) and again the average difference between the two groups of species is approximately twofold. Not only are the average differences in the predicted direction, but they are statistically highly significant ($P < 0.0001$) and they are consistent over all species within each group (see details in Supplementary Table S1). In addition to the consistency over species, there is also a consistency over all amino acids within the two codon groups. These differences are surprisingly large, given that we are dealing with a set of conserved orthologous proteins. Although taken as a group, the amino acid frequencies show approximately a twofold difference between the two groups (Fig. 2), at the level of individual amino acids, some of these differences are threefold or greater (Fig. 3). Again, the individual amino acid differences shown in Fig. 3 are highly statistically significant ($P < 0.0001$). Thus, we can conclude that strand asymmetries at the level of DNA have had a major influence on the composition of these mitochondrial protein sequences.

These results show that differences in the patterns of strand asymmetry between the coding and template strands of a mitochondrial gene can produce very significant changes in the amino acid composition of the encoded proteins. The magnitude of these changes is

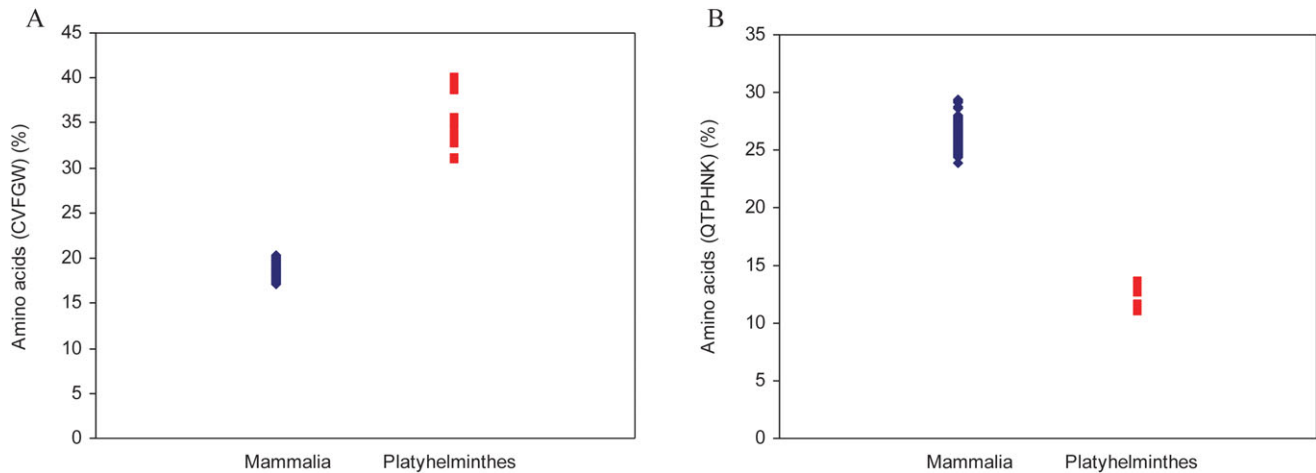


Figure 2. The amino composition of mitochondrial proteins reflects the DNA strand asymmetry. (A) The proportions of cysteine (C), Valine (V), phenylalanine (F), Glycine (G), and Tryptophan (W) are relatively low in mammalian proteins (shown in blue) and relatively high in flatworms (shown in red). These amino acids are encoded by GT-rich codons. (B) The proportions Glutamine (Q), Threonine (T), Proline (P), Histidine (H), Asparagine (N), and Lysine (K) are, in contrast, relatively high in mammals (shown in blue) and relatively low in flatworms (shown in red). These amino acids are encoded by CA-rich codons.

comparable to those noted previously by Foster et al.¹¹ for the effects of differences in mitochondrial nucleotide composition, i.e. GC content. It should be noted, however, that the strand asymmetries described in this study do not affect the same subsets of the amino acids as those affected by changes GC content. As an illustrative example, we can compare a single pair of mitochondrial genomes, one mammal and one flatworm, those of the red deer (*Cervus elaphus*) and the liver fluke (*Fasciola hepatica*). In both species, the overall nucleotide content of the mitochondrial coding sequences is virtually identical at 38% GC and proportions of GC-rich (GARP) and AT-rich (FYMINK) amino acids (see Foster et al.¹¹ for

details) are also very similar between these two species. In other words, their similarity in nucleotide composition is reflected in a similarity in the proportions of GC-rich and AT-rich amino acids. But when we compare the same two species for the levels of CA-rich (QTPHNK) and GT-rich (CVFGW) amino acids, we see large differences reflecting the differences in strand asymmetry between the mammals and the flatworms. For example, the liver fluke proteins contain more than twice as many Valine residues and more than five times as many cysteine residues as do their orthologs in the red deer. On the other hand, the liver flukes have approximately a third as many Glutamines and Threonines as are found in the red deer.

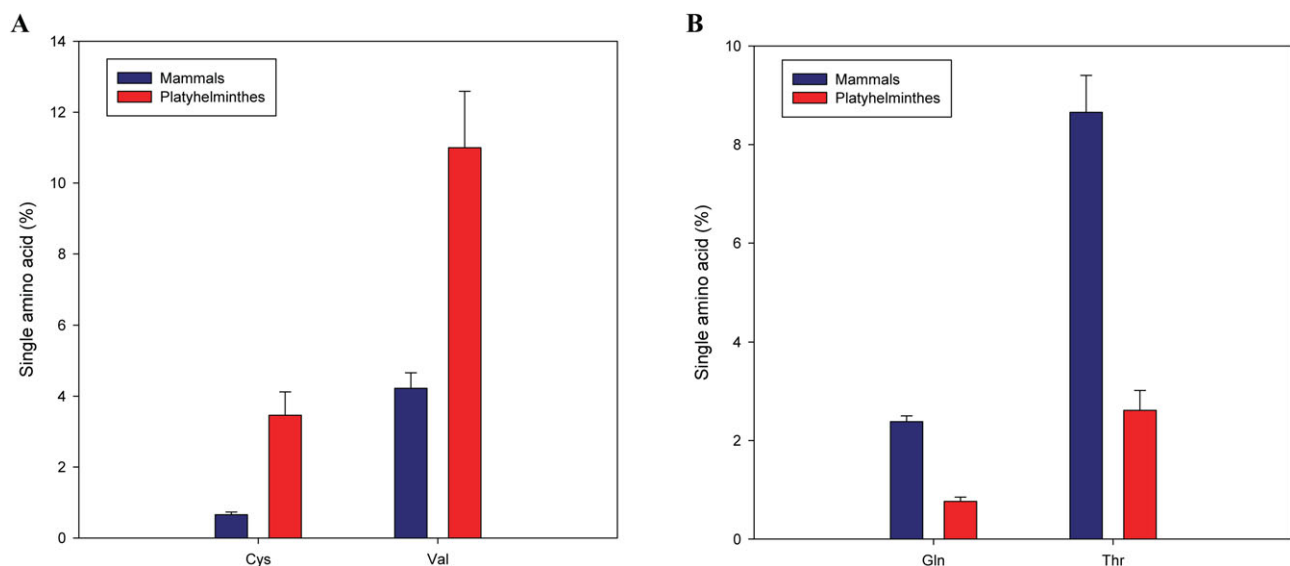


Figure 3. The proportions of individual amino acids that were most affected by DNA strand asymmetry in mammals and Platyhelminthes. (A) The proportions of cysteine and Valine are low in mammalian proteins (shown in blue) and high in flatworm proteins (shown in red). (B) The proportions of Glutamine and Threonine are high in mammals (shown in blue) and low in flatworms (shown in red).

These differences are also statistically highly significant ($P < 0.001$) and they are entirely consistent with what we see for the average differences between mammals and flatworms (Fig. 3 and Supplementary Table S1).

A possible alternative explanation for these results is that the amino acid differences are the cause, rather than the consequence, of the strand asymmetries. We can exclude this possibility, however, for two reasons. First, there is an even larger strand asymmetry at the synonymous codon sites (Supplementary Fig. S1) suggesting that the nucleotide skew is counterbalanced, to some extent, by functional constraint at the protein level. In other words, protein function does have an effect, but as a constraint rather than a cause. Another way to illustrate this point is to calculate the GC skew at each codon position separately. The results (Supplementary Table S2) show that the greatest differences in GC skew occur at the third codon position and the least skew occurs at the second position. This is consistent with the fact that many changes at the third codon position alter the codon usage but do not affect the protein sequence. Secondly, in mammalian mitochondria, one gene (*ND6*) is encoded on the opposite strand from the other 12 genes and, in accordance with our prediction, the amino acid composition of this mammalian protein displays a pattern that is similar to that of the flatworm proteins rather than the other 12 mammalian proteins (Supplementary Fig. S2). Both of these observations indicate that the primary effect is at the level nucleotide asymmetry between the two strands of the mitochondrial genome and that this DNA bias causes a secondary effect at the level of protein composition.

We performed a number of further tests to explore the interplay between functional constraint at the protein level and nucleotide skew at the DNA level. Since we limited our comparison to orthologous mitochondrial genes, we have eliminated the effect of different types of proteins in the two groups of species. Moreover, since mitochondrial function is highly conserved in metazoan animals, we expect that the orthologous proteins are performing essentially the same functions in the two species groups. There is still, however, the remote possibility that there are differences in physiological conditions between the mitochondria of mammals and those of flatworms and that these differences could contribute to the differences in protein composition which we observe between the two groups. To control for this possibility, we examined the sequences of three mitochondrial ribosomal protein genes: L11, L15, and L20 (Supplementary Table S2). Although these proteins function in the mitochondria, they are encoded by genes in the nuclear genome. The results show that the gene sequences do not show the characteristic mitochondrial gene skews and there is no significant difference between mammals and Platyhelminthes, either at the DNA level or at the

protein level. This indicates that the key factor underlying the species differences is not the functional environment of the proteins but rather the location of the genes encoding those proteins.

Since our results point to a mutational force at the DNA level that is counteracted by functional constraint at the protein level, we asked what would happen if we confined our analysis to a region of the protein where there was reason to believe that the functional constraint would be especially strong. The transmembrane domain portions of mitochondrial proteins provide an opportunity to do such an analysis. Specifically, we looked at the patterns of GC skew and amino acid composition in the transmembrane domain regions of the COXI gene. The results (Supplementary Table S2) show that the increased functional constraint on amino acid composition does indeed lead to a decreased difference in GC skew between the two species groups. For example, the difference in GC skew between mammals and flatworms for the complete dataset is 0.84, whereas it is reduced to 0.57 for the transmembrane domains, but the same pattern remains both at the DNA and protein levels. For instance, even within the transmembrane domains, the mammalian sequences have less cysteine and valine, and more glutamine and threonine than do their flatworm orthologs, reflecting the same patterns as shown in Figure 3. In other words, increased functional constraint decreases the effect of nucleotide skew, but it does not completely eliminate it.

Overall, our results show that the DNA strand asymmetry of animal mitochondrial genomes affects the amino acid composition of encoded proteins. A similar effect has been noted previously in bacterial genomes^{9,10} but it has not been reported in animal or plant genomes. Since we infer that the DNA strand bias is the cause rather than the consequence of the protein differences, this raises the question of what causes the DNA strand bias in the first place. In bacterial genomes, there is good evidence that it is related to DNA replication, based primarily on the fact that the direction of the bias switches at the origin of replication.^{4,7,8} Recent work indicates that although DNA strand biases are widespread in prokaryotes, eukaryotes, and viruses, the magnitude and the direction of the bias is variable, suggesting that the underlying causes are multifactorial.^{14,15} In animal mitochondrial genomes, the strand bias appears to be caused by varying durations of time that the heavy strand spends in the mutagenic single-strand state during replication.¹⁶ In addition to the replication-associated effects, there is evidence that transcription can also generate DNA strand asymmetry in eukaryotes due to transcription-coupled mutations.^{17,18}

A recent study has shown that strand bias in mitochondrial sequences can lead to artefactual results in phylogenetic reconstructions.¹⁹ This misleading result can be minimized by a recoding scheme that excludes transitions

at rapidly evolving neutral sites.^{19,20} Our results, however, show that mitochondrial strand biases can also have significant effects at non-neutral sites that change the amino acid sequence. This means that some degree of bias remains even after the correction has been implemented. As stated by Jones et al. (2007)²⁰ 'In the case of mitochondrial genes, strand-bias should be of particular concern and the previous use of mitochondrial genomes in resolving deep phylogenies requires critical re-evaluation'.

It remains to be seen if DNA strand asymmetry can also affect the composition of proteins encoded by eukaryotic nuclear genes, as has been shown for biases in the GC content of nuclear genes.^{21,22} It is already known that DNA strand asymmetries exist in nuclear genes^{17,18,23,24} but their effects on the composition of nuclear-encoded proteins have not been studied.

Supplementary Data: Supplementary data are available online at <http://dnaresearch.oxfordjournals.org>.

Funding

This work was supported by a grant to DAH from the Natural Science and Engineering Research Council of Canada.

References

1. Sharp, P. M., Averof, M., Lloyd, A. T., Matassi, G. and Peden, J. F. 1995, DNA sequence evolution: the sounds of silence, *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **349**, 241–247.
2. Lobry, J. R., 1997, Influence of genomic G + C content on average amino-acid composition of proteins from 59 bacterial species, *Gene*, **205**, 309–316.
3. Singer, G. A. and Hickey, D. A. 2000, Nucleotide bias causes a genomewide bias in the amino acid composition of proteins, *Mol. Biol. Evol.*, **17**, 1581–1588.
4. Lobry, J. R. 1996, Asymmetric substitution patterns in the two DNA strands of bacteria, *Mol. Biol. Evol.*, **13**, 660–665.
5. Sueoka, N. 1999, Two aspects of DNA base composition: G + C content and translation-coupled deviation from intra-strand rule of A = T and G = C, *J. Mol. Evol.*, **49**, 49–62.
6. Perna, N. T. and Kocher, T. D. 1995, Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes, *J. Mol. Evol.*, **41**, 353–358.
7. Mrázek, J. and Karlin, S. 1998, Strand compositional asymmetry in bacterial and large viral genomes, *Proc. Natl. Acad. Sci. USA*, **95**, 3720–3725.
8. McLean, M. J., Wolfe, K. H. and Devine, K. M. 1998, Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes, *J. Mol. Evol.*, **47**, 691–696.
9. Rocha, E. P. C., Danchin, A. and Viari, A. 1999, Universal replication biases in bacteria, *Mol. Microbiol.*, **32**, 11–16.
10. Lafay, B., Lloyd, A. T., McLean, M. J., Devine, K. M., Sharp, P. M. and Wolfe, K. H. 1999, Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases, *Nucleic Acids Res.*, **27**, 1642–1649.
11. Foster, P. G., Jermin, L. S. and Hickey, D. A. 1997, Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria, *J. Mol. Evol.*, **44**, 282–288.
12. Le, T. H., McManus, D.P. and Blair, D. 2004, Codon usage and bias in mitochondrial genomes of parasitic platyhelminthes, *Korean J. Parasitol.*, **42**, 159–167.
13. McManus, D. P., Le, T. H. and Blair, D. 2004, Genomics of parasitic flatworms, *Int. J. Parasitol.*, **34**, 153–158.
14. Rocha, E. P. and Danchin, A. 2001, Ongoing evolution of strand composition in bacterial genomes, *Mol. Biol. Evol.*, **18**, 1789–1799.
15. Rocha, E. P., Touchon, M. and Feil, E. J. 2006, Similar compositional biases are caused by very different mutational effects, *Genome Res.*, **16**, 1537–1547.
16. Faith, J. J. and Pollock, D. D. 2003, Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes, *Genetics*, **165**, 735–745.
17. Niu, D. K., Lin, K. and Zhang, D.Y. 2003, Strand compositional asymmetries of nuclear DNA in eukaryotes, *J. Mol. Evol.*, **57**, 325–334.
18. Touchon, M., Arneodo, A., d'Aubenton-Carafa, Y. and Thermes, C. 2004, Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes, *Nucleic Acids Res.*, **32**, 4969–4978.
19. Hassanin, A. 2006, Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution, *Mol. Phylogenet. Evol.*, **38**, 100–116.
20. Jones, M., Gantenbein, B., Fet, V. and Blaxter, M. 2007, The effect of model choice on phylogenetic inference using mitochondrial sequence data: lessons from the scorpions, *Mol. Phylogenet. Evol.*, **43**, 583–595.
21. Knight, R. D., Freeland, S. J. and Landweber, L. F. 2001, A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes, *Genome Biol.*, **2**, RESEARCH0010.
22. Wang, H. C., Singer, G. A. and Hickey, D. A. 2004, Mutational bias affects protein evolution in flowering plants, *Mol. Biol. Evol.*, **21**, 90–96.
23. Touchon, M., Nicolay, S., Arneodo, A., d'Aubenton-Carafa, Y. and Thermes, C. 2003, Transcription-coupled TA and GC strand asymmetries in the human genome, *FEBS Lett.*, **555**, 579–582.
24. Touchon, M., Nicolay, S., Audit, B., Brodie, E. B., d'Aubenton-Carafa, Y., Arneodo, A. and Thermes, C. 2005, Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins, *Proc. Natl. Acad. Sci. USA*, **102**, 9836–9841.