Asymmetrical Directional Mutation Pressure in the Mitochondrial Genome of Mammals

Aurelio Reyes,* Carmela Gissi,† Graziano Pesole,‡ and Cecilia Saccone*†

*Centro di Studio sui Mitocondri e Metabolismo Energetico, Consiglio Nazionale delle Ricerche, Bari, Italy; †Department of Biochemistry and Molecular Biology, University of Bari, Bari, Italy; and ‡Department of Biology, Defense and Agro-Forestal Biotechnology, University of Basilicata, Potenza, Italy

The base composition of 25 complete mammalian mitochondrial (mt) genomes has been analyzed taking into account all three codon positions (P123) and fourfold degenerate sites (P4FD) of H-strand genes. In the nontranscribed L strand, G is the less represented base and A is the most represented one in all cases, while C and T differ among species. H-strand protein-coding genes show an asymmetric distribution of the four bases between the two strands. The asymmetry indexes AT and GC skews on P4FD are much higher than those on P123, suggesting the existence of asymmetrical directional mutation pressure. Relationships between the compositional features and transcription or replication processes have been investigated in order to find a possible mechanism that could explain the origin of this asymmetry. AT and GC skews, the base composition in fourfold degenerate sites, and the number of variable sites for each gene are significantly correlated with the duration of single-stranded state of the H-stranded genes during replication. We tested different replication-related hypotheses, such as the existence of biased dNTP pools, γ DNA polymerase mispairing, and the asymmetric replication itself. Most of them failed to explain the observed results, hydrolytic deaminations being the only one in agreement with our data. Thus, we hypothesize that one of the crucial processes for the origin of asymmetric and biased base composition of mammalian mitochondrial genomes is the spontaneous deamination of C and A in the H strand during replication.

Introduction

Mammalian mitochondrial (mt) DNA is a very compact circular DNA molecule about 16.5 kb long, coding for 13 proteins which play a crucial role in electron transport and oxidative phosphorylation. It also encodes a complete set of 22 tRNAs and 2 rRNAs involved in the mitochondrial translation apparatus. Genes are asymmetrically distributed between the two strands, with one strand coding for 12 proteins, 14 tRNAs, and 2 rRNAs and the other coding for only one protein and 8 tRNAs (fig. 1).

The most remarkable feature of mammalian mt genomes, whose GC content ranges from 32.6% to 45.6%, is the uneven distribution of G and C between the two strands. The G-rich and G-poor strands, showing different buoyant density in a CsCl gradient, are called heavy (H) and light (L) strands, respectively. The transcribed strand (H) is more G-rich than the nontranscribed strand (L) and, thus, its mRNAs are very poor in G, particularly in the third codon positions. This results in a violation of parity rule type 2 (PR2) (Sueoka 1995), which refers to the intrastrand base composition expected at equilibrium, i.e., A = T and G = C, regardless of the GC content.

It is widely acknowledged that mutation rate is remarkably high in mtDNA. Indeed, a clearly higher nucleotide substitution rate is seen at synonymous codon positions, whereas the substitution rate of nonsynonymous positions varies from gene to gene and, in some cases, is lower for mitochondrial than for nuclear genes (Saccone, Pesole, and Kadenbach 1991). The higher mutation rate observed in mitochondria may be associated with oxidative damage produced by free radicals generated during the transport of electrons to oxygen that takes place at the level of the respiratory chain complex in the inner mitochondrial membrane. Other suggested causes include a lower proofreading activity of γ DNA polymerase during replication and the lack of DNA repair systems.

The replication of the mammalian mt genome is also asymmetric, with the two strands being synthesised from two distinct replication origins (fig. 1). The H-strand replication origin (OH) is located in the main noncoding region of the mtDNA, also called the D-loop region. The D loop is a triple-stranded structure produced by the nascent H strand, displacing the parental H strand. MtDNA replication starts with the elongation of the nascent H strand, expanding the D loop. When the displacement exposes the L strand replication origin (OL) as a single-stranded template about 11 kb downstream of the OH, the synthesis of the L strand starts in the opposite direction. During replication, the parental H strand remains single-stranded until paired by the newly synthesised L strand. It should be noted that, unlike the nuclear genome, mtDNA is not shielded by histone proteins. The parental H strand is only coated by single-stranded-mtDNA-binding proteins (mtSSBs) during replication.

The asymmetric mechanism of mtDNA replication could account for the strong compositional asymmetry observed in mitochondrial genomes. Since mtDNA replication is very slow, requiring about 2 h (Clayton 1982), the parental H strand remains single-stranded for a long time and is only partially protected by mtSSB proteins. During this time, the H strand is preferentially exposed to hydrolytic and oxidative damage and, therefore, is prone to mutations (Brown and Simpson 1982).
This introduces a mutation bias between the two strands generating the observed compositional asymmetry and the subsequent directional mutation pressure. The theory of directional mutation pressure, first postulated by Sueoka (1962), explains directional changes of compositional features due to a nonselective process (Sueoka’s two-parameter model). A more general model with no strand bias (six-parameter model) provides further theoretical background for analyzing directional mutation pressure (Sueoka 1995).

To study the compositional properties of mitochondrial genomes in relation to the mutation pattern and to test the existence of directional mutation pressure, we carried out an extensive compositional analysis of 25 complete genomes from 10 mammalian orders. In particular, for each of the 12 H-strand protein-coding genes, we investigated the relationship between the compositional features of all three codon positions $P_{123}$ and the third positions of fourfold degenerate codons $P_{4FD}$, i.e., glycine-GCN, valine-GTN, arginine-CGN, threonine-ACN, alanine-GCN, proline-CCN, leucine-CTN, and serine-TCN.

In each genome, the base composition of the 12 H-strand protein-coding genes was determined on the L-strand by the CODONTREE algorithm (Pesole, Attimonelli, and Saccone 1996). The percentage of each base was calculated on the whole gene ($P_{123}$) and the third positions of fourfold degenerate codons ($P_{4FD}$), i.e., glycine-GCN, valine-GTN, arginine-CGN, threonine-ACN, alanine-GCN, proline-CCN, leucine-CTN, and serine-TCN.

The GC and AT skews, which indicate compositional differences between the two strands, have been calculated according to the formulae by Perna and Kocher (1995):

$$
\text{GC skew} = \frac{G - C}{G + C}
$$

$$
\text{AT skew} = \frac{A - T}{A + T}
$$

where C, G, A, and T are the occurrences of the four bases in a given position of the gene ($P_{123}$ or $P_{4FD}$). The corresponding standard deviations (SDs) were calculated with Lobry’s (1996) formulae:

**Materials and Methods**

Compositional analyses were carried out on all 25 complete available mammalian mtDNAs, except for the *Mus domesticus* mtDNA because it is almost identical to the *Mus musculus* mtDNA. The complete list of the species analyzed, with the accession numbers (EMBL, release 51) of the corresponding mitochondrial sequences, is reported in table 1.

In each genome, the base composition of the 12 H-strand protein-coding genes was determined on the L-strand by the CODONTREE algorithm (Pesole, Attimonelli, and Saccone 1996). The percentage of each base was calculated on the whole gene ($P_{123}$) and the third positions of fourfold degenerate codons ($P_{4FD}$), i.e., glycine-GCN, valine-GTN, arginine-CGN, threonine-ACN, alanine-GCN, proline-CCN, leucine-CTN, and serine-TCN.

The GC and AT skews, which indicate compositional differences between the two strands, have been calculated according to the formulae by Perna and Kocher (1995):

$$
\text{GC skew} = \frac{G - C}{G + C}
$$

$$
\text{AT skew} = \frac{A - T}{A + T}
$$

where C, G, A, and T are the occurrences of the four bases in a given position of the gene ($P_{123}$ or $P_{4FD}$). The corresponding standard deviations (SDs) were calculated with Lobry’s (1996) formulae:

**Materials and Methods**

Compositional analyses were carried out on all 25 complete available mammalian mtDNAs, except for the *Mus domesticus* mtDNA because it is almost identical to the *Mus musculus* mtDNA. The complete list of the species analyzed, with the accession numbers (EMBL, release 51) of the corresponding mitochondrial sequences, is reported in table 1.

In each genome, the base composition of the 12 H-strand protein-coding genes was determined on the L-strand by the CODONTREE algorithm (Pesole, Attimonelli, and Saccone 1996). The percentage of each base was calculated on the whole gene ($P_{123}$) and the third positions of fourfold degenerate codons ($P_{4FD}$), i.e., glycine-GCN, valine-GTN, arginine-CGN, threonine-ACN, alanine-GCN, proline-CCN, leucine-CTN, and serine-TCN.

The GC and AT skews, which indicate compositional differences between the two strands, have been calculated according to the formulae by Perna and Kocher (1995):

$$
\text{GC skew} = \frac{G - C}{G + C}
$$

$$
\text{AT skew} = \frac{A - T}{A + T}
$$

where C, G, A, and T are the occurrences of the four bases in a given position of the gene ($P_{123}$ or $P_{4FD}$). The corresponding standard deviations (SDs) were calculated with Lobry’s (1996) formulae:
Thus, the compositional asymmetry between strands is stronger as it approaches 1 and is equal to 0 under no-strand-bias conditions (Sueoka 1995).

Sequence alignments and manual adjustments were made with the PILEUP and LINEUP programs (GCG 1994), respectively. The percentage of variable sites for each H-stranded protein-coding gene was calculated from the corresponding nucleotide multialignments.

Assuming the same rate of synthesis for both the H and L strands, the duration of the single-strand state of the parental H strand (D\textsubscript{ssH}) was set as the time passing between the arrival of the replicative fork O\textsubscript{H} and the arrival of the replicative fork O\textsubscript{L} in a specific region of the genome (fig. 1). According to this, the D\textsubscript{ssH} for each gene was calculated for the various genomes using the following formulae:

\[
D_{\text{ssH}} = \frac{L - 2(\bar{x} - O_L)}{L}
\]

for NDH1 and NDH2 genes and

\[
D_{\text{ssH}} = \frac{2(O_L - \bar{x})}{L}
\]

for the remaining genes, where L is the total length of the genome, O\textsubscript{L} is the position of the L strand origin of replication, and \(\bar{x}\) is the middle position of the gene. All distances were calculated following the replicative direction of the H strand, starting from the first position of the D-loop-containing region (fig. 1).

Mean D\textsubscript{ssH} values for each gene were also calculated. The relative order of the genes according to their mean D\textsubscript{ssH} values is as follows: COI < COII < ATP8 < ATP6 < COIII < ND3 < ND4L < ND4 < ND1 < ND5 < ND2 < Cytb.

Results
Base Composition and Asymmetry

Table 2 shows the base composition of the H-strand protein-coding genes of the mtDNAs on both P\textsubscript{123} and

<table>
<thead>
<tr>
<th>Latin Name</th>
<th>Common Name</th>
<th>Taxon</th>
<th>AC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Gorilla gorilla</td>
<td>Gorilla</td>
<td>Primates</td>
<td>D38114</td>
</tr>
<tr>
<td>2. Homo sapiens</td>
<td>Human</td>
<td>Primates</td>
<td>V00662</td>
</tr>
<tr>
<td>3. Hylobates lar</td>
<td>Gibbon</td>
<td>Primates</td>
<td>X99256</td>
</tr>
<tr>
<td>4. Pan paniscus</td>
<td>Pygmy chimpanzee</td>
<td>Primates</td>
<td>D38116</td>
</tr>
<tr>
<td>5. Pan troglodytes</td>
<td>Common chimpanzee</td>
<td>Primates</td>
<td>D38113</td>
</tr>
<tr>
<td>6. Pongo pygmaeus</td>
<td>Orang-utan</td>
<td>Primates</td>
<td>D38115</td>
</tr>
<tr>
<td>7. Bos taurus</td>
<td>Cow</td>
<td>Artiodactyla</td>
<td>V00654</td>
</tr>
<tr>
<td>8. Balaenoptera musculus</td>
<td>Blue whale</td>
<td>Cetacea</td>
<td>X72204</td>
</tr>
<tr>
<td>9. Balaenoptera physalus</td>
<td>Fin whale</td>
<td>Cetacea</td>
<td>X61145</td>
</tr>
<tr>
<td>10. Felis catus</td>
<td>Cat</td>
<td>Carnivora</td>
<td>U20753</td>
</tr>
<tr>
<td>11. Halichoerus grypus</td>
<td>Grey seal</td>
<td>Carnivora</td>
<td>X72004</td>
</tr>
<tr>
<td>12. Phoca vitulina</td>
<td>Harbor seal</td>
<td>Carnivora</td>
<td>X63726</td>
</tr>
<tr>
<td>13. Equus asinus</td>
<td>Donkey</td>
<td>Perissodactyla</td>
<td>X79337</td>
</tr>
<tr>
<td>14. Equus caballus</td>
<td>Horse</td>
<td>Perissodactyla</td>
<td>X79547</td>
</tr>
<tr>
<td>15. Rhinoceros unicornis</td>
<td>Indian rhinoceros</td>
<td>Perissodactyla</td>
<td>X97336</td>
</tr>
<tr>
<td>16. Cerotherium simum</td>
<td>White rhinoceros</td>
<td>Perissodactyla</td>
<td>Y07726</td>
</tr>
<tr>
<td>17. Oryctolagus cuniculus</td>
<td>Rabbit</td>
<td>Lagomorpha</td>
<td>AJ001588</td>
</tr>
<tr>
<td>18. Cavia porcellus</td>
<td>Guinea pig</td>
<td>Rodentia</td>
<td>AJ22276</td>
</tr>
<tr>
<td>19. Glis glis</td>
<td>Dormouse</td>
<td>Rodentia</td>
<td>AJ001562</td>
</tr>
<tr>
<td>20. Mus musculus</td>
<td>Mouse</td>
<td>Rodentia</td>
<td>V00711</td>
</tr>
<tr>
<td>21. Rattus norvegicus</td>
<td>Rat</td>
<td>Rodentia</td>
<td>X14848</td>
</tr>
<tr>
<td>22. Erinaceus europaeus</td>
<td>Hedgehog</td>
<td>Insectivora</td>
<td>X88989</td>
</tr>
<tr>
<td>23. Macropus robustus</td>
<td>Wallaroo</td>
<td>Metatheria</td>
<td>Y10524</td>
</tr>
<tr>
<td>24. Didelphis virginiana</td>
<td>Opossum</td>
<td>Metatheria</td>
<td>Z29573</td>
</tr>
<tr>
<td>25. Ornithorhyncus anatinus</td>
<td>Platypus</td>
<td>Prototheria</td>
<td>X83427</td>
</tr>
</tbody>
</table>

* Arnason and Gullberg (1993).
* Arnason et al. (1993).
* Xu, Gullberg, and Arnason (1996).
* Xu and Arnason (1994).
* Xu, Janke, and Arnason (1996).
* Xu and Arnason (1997).
* D’Erchia and Arnason (personal communication); D’Erchia et al. (1996).
* Janke, Xu, and Arnason (1997).
Table 2  
Base Composition and AT and GC Skew Regression Slopes

<table>
<thead>
<tr>
<th>LATIN NAME</th>
<th>LENGTH*</th>
<th>P123#</th>
<th>P4FD#</th>
<th>AT SLOPEd</th>
<th>GC SLOPEd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gorilla gorilla</td>
<td>10,845</td>
<td>29.38</td>
<td>32.38</td>
<td>11.88</td>
<td>26.36</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>10,845</td>
<td>29.29</td>
<td>33.15</td>
<td>12.01</td>
<td>25.36</td>
</tr>
<tr>
<td>Hylobates lar</td>
<td>10,845</td>
<td>29.14</td>
<td>33.55</td>
<td>12.48</td>
<td>24.84</td>
</tr>
<tr>
<td>Pan paniscus</td>
<td>10,845</td>
<td>29.87</td>
<td>32.47</td>
<td>11.53</td>
<td>26.23</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>10,845</td>
<td>29.50</td>
<td>32.55</td>
<td>11.78</td>
<td>26.18</td>
</tr>
<tr>
<td>Pongo pygmaeus</td>
<td>10,845</td>
<td>29.22</td>
<td>34.55</td>
<td>11.73</td>
<td>24.50</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>10,848</td>
<td>31.90</td>
<td>27.65</td>
<td>12.40</td>
<td>28.05</td>
</tr>
<tr>
<td>Balaenoptera musculus</td>
<td>10,845</td>
<td>31.29</td>
<td>29.93</td>
<td>11.76</td>
<td>27.03</td>
</tr>
<tr>
<td>Balaenoptera physalus</td>
<td>10,845</td>
<td>31.21</td>
<td>29.35</td>
<td>12.05</td>
<td>27.39</td>
</tr>
<tr>
<td>Felis catus</td>
<td>10,851</td>
<td>30.96</td>
<td>27.83</td>
<td>13.15</td>
<td>28.06</td>
</tr>
<tr>
<td>Halichoerus grypus</td>
<td>10,860</td>
<td>31.18</td>
<td>29.29</td>
<td>13.24</td>
<td>26.29</td>
</tr>
<tr>
<td>Phoca vitulina</td>
<td>10,860</td>
<td>31.20</td>
<td>29.27</td>
<td>13.27</td>
<td>26.26</td>
</tr>
<tr>
<td>Equus asinus</td>
<td>10,848</td>
<td>30.72</td>
<td>30.46</td>
<td>12.29</td>
<td>26.53</td>
</tr>
<tr>
<td>Equus caballus</td>
<td>10,842</td>
<td>30.76</td>
<td>30.38</td>
<td>12.36</td>
<td>26.50</td>
</tr>
<tr>
<td>Rhinoceros unicornis</td>
<td>10,851</td>
<td>32.05</td>
<td>29.33</td>
<td>11.75</td>
<td>26.86</td>
</tr>
<tr>
<td>Ceratotherium simum</td>
<td>10,854</td>
<td>31.64</td>
<td>29.68</td>
<td>12.09</td>
<td>26.60</td>
</tr>
<tr>
<td>Oryctolagus cuniculus</td>
<td>10,839</td>
<td>29.65</td>
<td>27.99</td>
<td>12.61</td>
<td>29.76</td>
</tr>
<tr>
<td>Cavia porcellus</td>
<td>10,842</td>
<td>30.57</td>
<td>26.06</td>
<td>13.12</td>
<td>30.26</td>
</tr>
<tr>
<td>Glis glis</td>
<td>10,836</td>
<td>30.60</td>
<td>24.61</td>
<td>11.83</td>
<td>32.95</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>10,839</td>
<td>33.08</td>
<td>26.11</td>
<td>11.46</td>
<td>29.35</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>10,860</td>
<td>32.38</td>
<td>28.35</td>
<td>11.34</td>
<td>27.92</td>
</tr>
<tr>
<td>Erinaceus europaeus</td>
<td>10,839</td>
<td>31.79</td>
<td>20.95</td>
<td>11.54</td>
<td>35.71</td>
</tr>
<tr>
<td>Macropus robustus</td>
<td>10,854</td>
<td>30.73</td>
<td>28.58</td>
<td>12.08</td>
<td>28.62</td>
</tr>
<tr>
<td>Didelphis virginiana</td>
<td>10,983</td>
<td>33.17</td>
<td>22.80</td>
<td>11.08</td>
<td>32.95</td>
</tr>
<tr>
<td>Ornithorynchus anatinus</td>
<td>10,860</td>
<td>30.93</td>
<td>25.69</td>
<td>12.51</td>
<td>32.47</td>
</tr>
<tr>
<td>Overall</td>
<td>30.82</td>
<td>28.92</td>
<td>12.13</td>
<td>28.13</td>
<td>46.46</td>
</tr>
<tr>
<td>SD</td>
<td>1.20</td>
<td>3.32</td>
<td>0.61</td>
<td>2.82</td>
<td>5.08</td>
</tr>
</tbody>
</table>

# Length of the H strand protein-coding genes.
# All three codon positions.
# Slope value of the regression line of AT skew on P4FD against the single-stranded state duration.
# Slope value of the regression line of GC skew on P123 against the single-stranded state duration.

P4FD- Mammalian species show a rather heterogeneous base composition, which is much more evident on P4FD. The only common feature is the low G%. In all species, A is the most represented base, but in Primates, a high C%, comparable with the A%, is observed on both P123 and P4FD. On P123, C and T are equally represented in most mammalian orders, except in Primates, where C% > T%, and in the dormouse, hedgehog, opossum, and platypus, where C% < T%. It is noteworthy that this opposite pattern of C/T occurrence is found mainly in species located at the basal branch of the mammalian tree (Reyes, Pesole, and Saccone 1998). It is also remarkable that A% and G% present a significantly lower variability (on average 1.6 and 0.6% of the mean value, respectively) than C% and T% (3.3% and 2.8%, respectively). On P4FD, because of the high variability of G% and C%, we do not detect a common pattern of pyrimidine representation among the different species.

The GC and AT skews on P123 and P4FD have been calculated for each mitochondrial genome as a measure of the compositional asymmetry. The GC skew is negative and the AT skew is positive in all cases except for the rabbit, dormouse, hedgehog, and platypus, for which P123 AT skew is negative. Moreover, absolute values of GC skew are always higher than those for AT skew, except in the two seals, where the two P4FD skews are similar. Figure 2 shows the plot of both AT and GC skews on P4FD on the corresponding skews in P123 of each genome. In all species, the absolute values of the skews on P4FD are higher than those on P123, and the standard deviations are very low. Indeed, mean and SD of GC and AT skews are between −0.41 ± 0.05 and 0.45 ± 0.12 on P123, whereas −0.73 ± 0.08 on P4FD, respectively. Thus, the strand asymmetric base composition is stronger in weakly constrained sites. Because all species present similar skew values, especially in the case of AT skew, dots are centered within a rather small range of variation (fig. 2). Moreover, there is no relationship between phylogenetic position and asymmetric base composition, as closely related species do not plot together in the graph. Thus, for example, the Primates do not appear close to one another but are dispersed over the graph (fig. 2).

Compositional Variation from O₃

In order to study the possible correlation between mtDNA compositional asymmetry and the single-stranded state of the H strand during replication, we plotted the AT and GC skews calculated for each gene on both P123 and P4FD against the Dₘₜₐₜ of the corresponding gene (fig. 3). In P123, no correlation was found between AT skew and Dₘₜₐₜ, while a negative correlation was detected for the GC skew. On P4FD, both AT and GC skews show a very significant correlation, with the
Fig. 2.—Plot of the AT and GC skew on the third positions of fourfold degenerate codons ($P_{4FD}$) against the corresponding skew on all three codon positions ($P_{123}$) for each mtDNA analyzed. For GC skews, absolute values are presented. Species are numbered according to table 1. Standard deviations were calculated according to Lobry (1996).

AT skew positively correlated ($r = 0.85, P < 0.005$) and the GC skew negatively correlated ($r = -0.74, P < 0.01$) with the single-stranded state. In general, in the less constrained positions ($P_{4FD}$), compositional asymmetry increases significantly with the duration of the single-stranded state during replication. The same trend is observed when all the mammalian species are considered individually. Table 2 gives the slopes of $P_{4FD}$ AT and GC skew regression lines on $D_{ssH}$ for each species. In all cases, the AT skew increases and the GC decreases with $D_{ssH}$, but the slope appears to be species-specific. Thus, we find that closely related species may have very different slopes, as in the case of the two chimpanzees or within the order Primates. This would sug-

Fig. 3.—Correlation between the AT and GC skews calculated on all three codon positions ($P_{123}$) and skew on the third positions of fourfold degenerate codons ($P_{4FD}$) of each H-stranded gene and its duration as a single strand ($D_{ssH}$) ($** P < 0.01; *** P < 0.005$).
Fig. 4.—Correlation between the nucleotide percentages at the third positions of fourfold degenerate codons ($P_{4FD}$) and the single-stranded state duration ($D_{ssH}$) for each H-stranded gene (*$P < 0.05$; ***$P < 0.005$).

Fig. 5.—Plot of the percentage of nucleotide variable sites against the single-stranded state duration ($D_{ssH}$) for each H-stranded gene. The equation and correlation value refer to the data without ATP8 and Cytb (***$P < 0.01$).

suggest that the time required for replication of the mtDNA may be variable among species.

The possible correlation between the base composition on $P_{4FD}$ of the H-strand protein-coding genes and $D_{ssH}$ was investigated (fig. 4). All four bases exhibit a statistically significant correlation with $D_{ssH}$. In general, the A and C frequencies increase with the $D_{ssH}$, while the G and T frequencies decrease. A and G have about the same slope and correlation coefficient, although in opposite directions; the same results are obtained for C and T.

To investigate a possible correlation between a gene's location and its degree of functional constraints, we studied the correlation between the number of variable sites for each gene and $D_{ssH}$ (fig. 5). A positive correlation ($r = 0.34, P > 0.05$) was found, although it is not significant due to the deviation of the ATP8 and Cytb genes. Indeed, when these two genes are omitted, a significant positive correlation is found ($r = 0.83, P < 0.01$).

Discussion

Despite the relatively high degree of variability in the base composition of mammalian genomes, G was the least represented base in both $P_{123}$ and $P_{4FD}$, while A was the most represented one, mainly in $P_{123}$ (table 2). Moreover, there is a clear violation of PR2 (Sueoka 1995); i.e., the intrastrand equalities $A = T$ and $G = C$ expected at equilibrium are not obeyed (table 2). Deviations from PR2 can be explained by the fact that equilibrium has not yet been reached or by the existence of an asymmetric substitution matrix. In the present case, because of the high substitution rate of the mtDNA and the similar behavior observed on $P_{123}$ in all species, it
can be postulated that equilibrium, or a close-to-equilibrium state, has been reached. Thus, the lack of intrastrand PR2 may be due to an asymmetric substitution matrix. The existence of lower absolute values of AT than GC skews (fig. 2) can be explained by unequal substitution rates, at least between α-base (A or T) and γ-base (G and C). Previous results based on 4 mitochondrial genomes (Sacccone et al. 1990) and 43 human individuals (Tanaka and Ozawa 1994) also confirmed the existence of an asymmetric substitution matrix. The existence of this matrix would suggest a directional mutation pressure or the action of selection on mtDNA (Sueoka 1995). Even though the deviation from PR2 is observed on both P_{123} and P_{4FD} (table 2 and fig. 2), the highest asymmetry values are detected on the less constrained sites, strongly supporting an asymmetrical directional mutation pressure, rather than selection, as an explanation of the asymmetrical base composition of mtDNAs.

The existence of directional mutation pressure has been widely reported in bacteria (for a review see Francino and Ochman 1997), but it has not previously been described in mammalian mtDNAs, although indirect evidence can be drawn from previous surveys (Jermiin, Graur, and Crozier 1995).

The underlying mechanism that leads to the observed compositional bias in the mammalian mtDNA could be related to transcription or replication (Francino and Ochman 1997). In the case of replication, different hypotheses have been postulated: (1) a bias in the intramitochondrial dNTP pools during replication, (2) a strand-specific mispairing rate of the γ DNA polymerase, or (3) a consequence of the replication process itself that is known to be asymmetric (Parsons and Simpson 1973; Clayton 1982; Sacccone and Sbisa 1994).

Transcription induces a transient single-stranded state of the nontranscribed DNA strand; thus, the mutation pattern between transcribed and nontranscribed strands might be biased by a transcription-coupled repair system or by deamination (Francino and Ochman 1997). Transcription-coupled repair has been experimentally shown to produce an excess of C-to-T mutations on the nontranscribed strand that leads to G-to-A substitutions on the transcribed strand (i.e., a positive GC and AT skew) and to a positive correlation between the level of gene expression and strand-asymmetry intensity. An increase of C deamination, and consequently of C-to-T mutations, in the nontranscribed strand has also been observed in actively expressed Escherichia coli genes (Beletskii and Bhagwat 1996). In mtDNA, we found a completely different pattern of strand asymmetry: GC skews are negative and AT skews are positive. Moreover, genes encoded by the same strand are equally expressed in a polycistronic transcript (Clayton 1984), but they have different skew values depending on the D_{ah}.

A base composition biased toward the transcribed strand can be also due to the relative abundance of tRNA, which affects the synonymous codon usage frequency. In mtDNA, although there is a strong biased codon usage (Sacccone and Sbisa 1994), no data are available on tRNA abundance.

The low G% of the nontranscribed strand and then in the transcription product might also be explained as a protection mechanism of transcripts from oxygen radical damage whose major base lesion is G-directed and produces 8-hydroxyguanine (Ames, Shigenaga, and Hagen 1995).

Regarding replication-related hypotheses, a biased pool of dNTPs would result in a biased base composition. However, this would hardly explain the strong asymmetry between the strands (fig. 3). In vitro experiments carried out on bacteria have shown that an excess of a single dNTP substrate over the other three generates well-defined strand-specific errors (Roberts, Thomas, and Kunkel 1991). However, the presence of biased dNTP pools in mitochondria has not yet been established.

Concerning γ DNA polymerase mispairing, empirical data indicate low proofreading efficiency (Kunkel and Alexander 1986; Kunkel and Soni 1988; Pinz, Shibutani, and Bogenhagen 1995) as well as different degrees of mispairing for each nucleotide. The mispairing of A with C is 23-fold more frequent than that of G with T (Kunkel and Alexander 1986). Whether these two properties of the γ DNA polymerase contribute to the observed compositional bias cannot be settled with our data.

Both biased dNTP pools and γ DNA polymerase would actually explain a biased base composition but never an asymmetric distribution. An alternative hypothesis to explain the peculiar compositional bias in mitochondrial DNA takes into account the fact that, during its replication, mtDNA remains single-stranded for two thirds of the replication cycle. This asymmetry in the replication prevents the parental L strand from being single-stranded in any phase of the replication, while genes from the parental H strand remain single-stranded for variable periods of time, up to nearly 80 min in the case of cytochrome b (Clayton 1982). According to this hypothesis, the longer the H strand remains single-stranded, the higher the probability of it suffering from spontaneous decomposition by nonenzymatic methylations, oxidation, and hydrolysis.

A significant correlation between compositional features of H-strand encoded genes and the duration of their single-stranded state during replication (D_{ah}) would support the hypothesis above. Indeed, a significant correlation is observed between base frequencies on fourfold degenerate sites (P_{4FD}) and D_{ah}, with A and C frequencies increasing and G and T frequencies decreasing (fig. 4). The increase in C frequency rate is similar to the decrease in T frequency rate, and the same is observed for A and G frequencies.

Methylations do not explain the observed variation, as they do not modify base-pairing or are cytotoxic. Similarly, neither is this pattern explained by the increased formation of 8-hydroxyguanine by oxidation (Lindahl 1993), which preferentially pairs with adenine rather than with cytosine and would generate some correlation between L-strand A% increase and C% decrease. On the contrary, we observe that both A and C increase as a function of D_{ah}.
Fig. 6.—Schematic representation of the suggested deamination processes that may take place in the single-stranded H strand: cytosine (C) into uracil (U) and adenine (A) into hypoxanthine (hX), which would imply a change of G into A and of T into C on the L strand.

The hydrolytic deaminations of both C and A depending on the single-stranded state might be responsible for the observed correlation between base frequencies and $D_{\text{shift}}$. The mutation pattern we hypothesize is summarized in figure 6, showing how the deamination of both cytosine and adenine supposedly plays a major role. Spontaneous deamination of cytosine on the H strand produces uracil which base-pairs with A rather than G, and consequently the percentage of G decreases and that of A increases on the L strand according to $D_{\text{shift}}$. In the same way, the deamination of A on the H strand results in hypoxanthine (Lindahl 1993), which base-pairs with C rather than T, resulting in a reduction, correlated to $D_{\text{shift}}$, in the T percentage and an increase in C on the L strand.

All these changes are detectable in P4FD (fig. 4), which are positions under no or limited selection. When intergenic regions of mammalian mitochondrial DNA, also considered selectively neutral, are analyzed (Jermini, Graur, and Crozier 1995), a strand-specific bias toward C and A is observed. In the present case, the same trend is also detected in all three codon positions (fig. 3), even though it is not significant due to constraints on the gene product. In agreement with these results, other authors have reported that mutations in human mitochondrial DNA are highly asymmetric, and have suggested that there are distinct mutational gradients for each of the nucleotides on P4FD with respect to O$_L$, which could be partially explained by deamination (Pepe et al. 1983; Saccone et al. 1993; Tanaka and Ozawa 1994).

The strand displacement mechanism of mtDNA replication involves mtSSBs (Mignotte, Barat, and Mounolou 1985; Van Tuyle and Pavco 1985), which might not be able to completely protect the DNA from deamination. In the case of nuclear DNAs, where replicative intermediates are also protein-coated, single-strand deamination of cytosine has been found to be 140- to 200-fold higher than that in double-stranded DNA (Sancar and Sancar 1988; Frederico, Kunkel, and Shaw 1990). Moreover, cytosine deamination in *Saccharomyces cerevisiae* has been estimated to be 40-fold higher than that in *E. coli*, possibly due to a slower rate of eukaryotic transcription, which might keep DNA locally in single-stranded form for a longer period (Impellizzeri, Anderson, and Burgers 1991). Thus, there is a clear correlation between the time spent as a single strand and the possibility of suffering deamination, especially as far as cytosine is concerned.

Assuming that deamination is the only mechanism involved in determining the observed compositional bias, we calculated the relationship between deamination of cytosine and adenine, taking into account the mean values of each base in the mitochondrial genome (relationship between mean slope values for A-G or C-T divided by the frequency of deaminated base, $A_H$ and $C_H$, respectively). According to our results, the two deamination events do not appear to be equally likely, with deamination of cytosine being twofold higher than deamination of adenine. This rough estimate is in agreement with previous results, in which deamination of A was found to be slower than deamination of C in both nuclear (Lindahl 1993) and human mitochondrial DNA (Tanaka and Ozawa 1994).

If the most important factor affecting both compositional bias and rate of mutation is the length of time genes remain in the single-stranded state, more constrained genes should be located closer to O$_L$ than should less constrained genes. Since we have measured the degree of functional constraints as the number of variable nucleotide sites, genes remaining for less time as single-stranded should have a lower number of variable sites than those remaining longer. A positive correlation between these two factors has been found ($r = 0.340$, $P > 0.05$) and, when ATP8 and Cytb are not considered in the analysis, the correlation becomes highly significant ($r = 0.834$, $P < 0.01$; fig. 5). In the case of ATP8, we find a higher value of variable sites than expected, while in Cytb, lower values of variable sites are found. ATP8 is a short gene that seems to accept almost every kind of change, probably because it is subject to low selection in most domains of the molecule, and thus high levels of variability both at the amino acidic and nucleotidic levels are observed. Consequently, the analysis of vertebrate mitochondrial genomes (Kumar 1996) has revealed that ATP8 is the fastest-evolving mitochondrial gene, even at the first and second codon positions. In echinoderms, too, this gene evolves extremely fast, showing a high number of variable sites (De Giorgi et al. 1997). In contrast, a lower number of variable sites than expected from its position in the genome is observed for Cytb. Most of its sites are highly conserved in mammals, especially in the Qo and Qi regions (Irwin, Kocher, and Wilson 1991; Ma et al. 1993). Thus, this gene would be under strong selection, resulting in high variation only on synonymous codon positions.

In conclusion, we have found an asymmetrical base composition to be most likely due to the existence of an asymmetrical directional mutation pressure. According to our data, the underlying mechanism that better explains these results is the deamination of C and A on the H strand, which has been found to be correlated with...
the time genes remain single-stranded during replication. Nevertheless, the data so far available are not sufficient to reject other possible mechanisms.

Acknowledgments

We thank Drs. U. Arnason and A. M. D’Erchia for providing us with the positions of the genes within the mitochondrial genome of the guinea pig. The constructive comments of an anonymous referee are greatly appreciated. We also thank M. Lonigro and two anonymous reviewers for revising the English text. This study was supported by grants from MURST (Italy) and the Human Capital and Mobility European Project ERBCHR XCT 930254.

LITERATURE CITED


PINZ, K. G., S. SHIBUTANI, and D. F. BOGENHAGEN. 1995. Action of mitochondrial DNA polymerase gamma at sites of
base loss or oxidative damage. J. Biol. Chem. 270:9202–9206.


DAN GRAUR, reviewing editor

Accepted April 8, 1998