Patterns of Nucleotide Composition at Fourfold Degenerate Sites of Animal Mitochondrial Genomes

Nicole T. Perna, Thomas D. Kocher

Program in Genetics and Department of Zoology, University of New Hampshire, Durham, NH 03824, USA

Received: 14 February 1994 / Accepted: 6 January 1995

Abstract. Three statistics (%GC, GC-skew, and AT-skew) can be used to describe the overall patterns of nucleotide composition in DNA sequences. Fourfold degenerate third codon positions from 16 animal mitochondrial genomes were analyzed. The overall composition, as measured by %GC, varies from 3.6 %GC in the honeybee to 47.2 %GC in human mtDNA. Compositional differences between strands of the mitochondrial genome were quantified using the two skew statistics presented in this paper. Strand-specific distribution of bases varies among animal taxa independently of overall %GC. Compositional patterns reflect the substitution process. Description of these patterns may aid in the formation of hypotheses about substitutional mechanisms.

Key words: mtDNA — Composition — Skew — %GC

Introduction

The nucleotide composition of mitochondrial genomes varies among animal taxa. For example, the complete mitochondrial DNA sequence of the honeybee is only 15.1% GC (Crozier and Crozier 1993) while that of the human (Anderson et al. 1981) is 44.4% GC. Compositional differences also exist between the two strands of the mitochondrial genome and were originally recognized as differences in buoyant density in CsCl gradients (Brown 1981). The biochemical and evolutionary origins of these compositional features and the relationship between the strand-specific distribution and the overall %GC of the genome are presently unknown. Clearly, nucleotide usage results from the process of substitution, but many of the factors which affect the pattern and rate of substitution in mtDNA are not well characterized.

Underlying the process of substitution is a mutational spectrum created by misincorporation of nucleotides by DNA polymerases (Kunkel 1985) and spontaneous chemical degradation (Lindahl 1993). Base mismatches created by these factors may then be resolved by repair mechanisms or lead to mutations. Finally, this collection of mutations is filtered by selection for function at either the level of the DNA or the product it encodes. Thus, patterns in composition within genomes and compositional differences among homologous sequences could result from both variation in the selective constraints and changes in the mutational spectrum during evolutionary divergence.

In this paper we use three measures to describe nucleotide patterns at fourfold degenerate third codon positions because, of all the sites in the mitochondrial genome, these are most likely to reflect the underlying mutational matrix. In studies of nuclear genomes, non-coding or pseudogene sequences are often used to study the mutational matrix (for example, Balmer 1985). Unfortunately, the major noncoding region of mtDNA has important, if poorly understood, functions which exert strong selective constraint (Kocher and Wilson 1991). Likewise, patterns of selection on tRNA and rRNA genes arising from secondary structure and interactions with other molecules are complex, and it is difficult to define
a homogeneous subset of sites from these genes (Xiong and Kocher 1993). First and second codon positions are subject to selection for amino acid sequence in the resulting protein, and thus are not good estimators for compositional patterns generated by the mutational spectrum. Some studies of mtDNA composition (Asakawa et al. 1991) have included all third codon positions. Mitochondrial proteins have highly biased amino acid composition and unequal numbers of twofold degenerate codons may affect estimates of equilibrium composition.

Although fourfold degenerate sites are free of the selective constraints of amino acid specification, composition at these sites may still be affected by selection for translational efficiency. Synonymous codon usage in bacteria and yeast is strongly correlated with overall composition of the genome, yet highly expressed genes often use a higher proportion of certain codons to promote efficient translation (Shields and Sharp 1987). Selection to match codon usage with iso-accepting tRNA abundance is unlikely to be important in the mitochondrial system, where there is usually only one tRNA for each fourfold degenerate codon family. However, there might be selection among synonymous codons for different binding affinities to the tRNA antecedent (Bulmer 1991). Asakawa et al. (1991) argue that this is also unlikely to be a factor in mitochondrial composition because mitochondrial genome rearrangements have periodically caused some genes to switch strands. These genes evolve base compositions consistent with their new location rather than their original strand.

If composition of fourfold degenerate sites is primarily the result of the mutational matrix, then it is clear that the directional mutation pressure modeled by Sueoka (1988) to address variation in GC among eukaryotic nuclear and bacterial sequences. We hope that an accurate and quantitative description of compositional patterns at fourfold degenerate sites, and an exploration of the evolutionary history of compositional variation in mtDNA, will provide insight into the nature of the mutational pressures acting on this molecule.

Materials and Methods

Sequences. Complete mitochondrial genome sequences are available for 13 taxa included in this analysis: Apis mellifera (Crozier and Crozier 1993), Ascaris suum (Okimoto et al. 1992), Bos taurus (Anderson et al. 1982), Canis familiaris elegans (Okimoto et al. 1992), Crocodylus niloticus (Teng et al. 1992), Cyprinus carpio (Chung et al. 1994), Drosoophila yakuba (Clay and Whalenbehen 1985), Gallus gallus (Dejardin and Mounis 1990), Homo sapiens (Anderson et al. 1981), Mus musculus (Bibb et al. 1981), Paracronosaurus lividus (Cantatore et al. 1989), Peronsonus marinus (Lee and Kocher 1995), and Strongylocentrotus purpuratus (Jacobs et al. 1988). For these taxa, we used fourfold degenerate codon positions from all the mitochondrial protein coding genes which are encoded on one strand of the genome, known as the heavy strand in vertebrates (Brown 1981). The motivation for using only genes encoded on the same strand is to minimize the potential confounding compositional effects of different evolutionary pressures experienced by sequences transcribed at different rates (Antardt et al. 1982) and replicated asymmetrically (Clayton 1992). In nematodes, the replication mechanism is unknown and cannot be inferred by phylogenetic comparison, but all 12 protein coding genes are on the same strand and are included in this analysis. At the time of this analysis, complete mitochondrial genome sequence is not available for any mollusk, so we have filled this phylogenetic gap with sequences from 12 Mytilus edulis genes, all of which are encoded on the same strand (Hoffmann et al. 1992). We have also included data from partial genome sequences for two additional echinoderm taxa, Arbacia livida (De Giorgi et al. 1991a) and Asterinapectinifera (Hirano et al. 1987), because the replication mechanism differs between urchins and sea stars and a strand-specific mutation pattern has been observed. Arbacia (De Giorgi et al. 1991b), Partial ND-5 (185aa) and COII (109aa) sequences were used for A. livida. The A. pectinifera data came from partial COIII (69aa) and NDS (512aa) and complete ND3 and ND4 sequences.

Codon Frequencies. The composition of fourfold degenerate third codon positions of the 16 species was calculated by generating codon frequency tables with the GCG CODONP program (Devereux et al. 1984) and summing the frequency of each base at the third positions across the 8 fourfold degenerate codon families (guanine-GCN, leucine-CTN, valine-CTN, arginine-CGG, threonine-ACN, alanine-GCN, serine-TCN, and proline-CCN) common to all variations of the animal mitochondrial genetic code. The CN leucine codons were included even though they are actually sixfold degenerate. The additional two-fold degeneracy results from synonymous first codon positions. A large proportion of the total fourfold degenerate codons found in mitochondrial proteins are CTN leucines. We were concerned that first position substitutions in the twofold degenerate TTR leucine codons might influence the number of codons from the fourfold family which end in A and G. An analysis was performed on a partial data set and also without the inclusion of the leucine codon family. No significant differences were seen (data not shown).

Statistics. We calculated three measures of compositional distribution from the fourfold degenerate third codon position nucleotide frequency data. Complementary pairing of bases permits all three to be calculated from the frequencies of nucleotides on a single strand. The overall composition of the double-stranded molecule is measured by the proportion of G + C out of the total. This is a commonly used measure most frequently and simply described as %GC.

The other two measures describe the compositional difference between the two strands:

\[ GC\text{-SKEW} = (G - C)/(G + C) \]  
\[ AT\text{-SKEW} = (A - T)/(A + T) \]

where G, A, T, and C are the frequencies of each nucleotide from the sense strand.

These two equations differ from other measures of strand-specific composition in several ways. The bias statistic used by Irwin et al. (1991) measures the deviation of the four bases from equal frequency, confounding %GC with strand-specific compositional patterns. Thompson and Wilson (personal communication) have recommended the statistic:

\[ \text{Shew} = 2(G1 - C1 + A1 - T1) \]

where G1, C1, A1, and T1 are proportions of each nucleotide on a single strand of the DNA helix. The most problematic feature of this
Table 1. Nucleotide composition of fourfold-degenerate third codon positions from 16 animal taxa

<table>
<thead>
<tr>
<th>Genus</th>
<th>%G</th>
<th>%A</th>
<th>%T</th>
<th>%C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caenorhabditis elegans</td>
<td>8.2</td>
<td>37.5</td>
<td>50.2</td>
<td>4.1</td>
</tr>
<tr>
<td>Ascaris suum</td>
<td>13.4</td>
<td>7.1</td>
<td>77.6</td>
<td>2.7</td>
</tr>
<tr>
<td>Drosophila yakuba</td>
<td>2.2</td>
<td>49.5</td>
<td>46.1</td>
<td>2.3</td>
</tr>
<tr>
<td>Apis mellifera</td>
<td>0.2</td>
<td>66.7</td>
<td>29.7</td>
<td>3.4</td>
</tr>
<tr>
<td>Mytilus edulis</td>
<td>26.5</td>
<td>27.4</td>
<td>34.6</td>
<td>11.4</td>
</tr>
<tr>
<td>Potamonax marinus</td>
<td>2.7</td>
<td>44.2</td>
<td>31.8</td>
<td>21.3</td>
</tr>
<tr>
<td>Crithon carpio</td>
<td>5.7</td>
<td>53.0</td>
<td>14.2</td>
<td>27.1</td>
</tr>
<tr>
<td>Crenorina lanceolata</td>
<td>8.5</td>
<td>40.7</td>
<td>15.2</td>
<td>35.5</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>3.3</td>
<td>43.6</td>
<td>12.0</td>
<td>41.1</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>4.5</td>
<td>51.0</td>
<td>16.0</td>
<td>28.6</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>4.9</td>
<td>55.6</td>
<td>17.9</td>
<td>22.6</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>5.1</td>
<td>39.4</td>
<td>13.4</td>
<td>42.1</td>
</tr>
<tr>
<td>Sirengitoecius purpuratus</td>
<td>10.3</td>
<td>36.3</td>
<td>26.6</td>
<td>26.8</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>9.3</td>
<td>43.4</td>
<td>24.1</td>
<td>23.2</td>
</tr>
<tr>
<td>Arctica islandica</td>
<td>4.9</td>
<td>36.6</td>
<td>40.7</td>
<td>17.7</td>
</tr>
<tr>
<td>Asterina pectinifera</td>
<td>8.2</td>
<td>34.7</td>
<td>25.1</td>
<td>31.9</td>
</tr>
</tbody>
</table>

statistic is that it confounds two aspects of skew, that arising from AT pairs with that involving GC base pairs. Numerically equivalent values of this statistic can arise from very different patterns of skew. We feel that it is important to consider the contribution of each type of nucleotide pair to the overall skew. Also, it is usually most convenient to express such statistics within a range from 0 to 1. We have also chosen to standardize our skew measures by the composition of the double-stranded molecule in order to consider different, possibly independent, patterns of composition. Finally, we have eliminated the absolute-value bars to distinguish the direction of each type of skew. These measures are similar to those used by Saccor et al. (1993), differing only in removal of absolute-value bars in the numerator. It is important to note that the sign of the skew statistic is meaningful only with reference to a particular strand since the same values with the opposite sign describe the composition of the other strand. However, the sign will provide an additional level of discrimination in comparative analyses where we can unambiguously identify one strand with reference to an asymptotic biochemical process, like direction of replication.

Results and Discussion

Base composition for the fourfold degenerate sites from 16 taxa is reported in Table 1. These nucleotide frequencies were used to calculate the %GC and values for both skew measures found in Fig. 1. It is useful to consider these characteristics in a phylogenetic context. Examination of the phylogenetic distribution of compositional patterns allows the formation of hypotheses about which lineages have experienced changes in the substitution process. The phylogeny is derived from a combination of morphological and molecular evidence and represents a consensus of current opinion on the relationship of the taxa used in this study. (See Sidow and Thomas 1994.)

Nematodes

Although the two nematodes, C. elegans and A. suum, have a very similar overall composition, 12.3 and 16.1 %GC, there are dramatic differences in the way the ~85% A’s and T’s are distributed on the two strands of each genome. In C. elegans, the sense strand has 37.5% A and 50.2% T, whereas in Ascaris, the same strand has only 7.1% A and 77.6% T. Ascaris has a much stronger AT-skew than C. elegans. The GC-skew of these two taxa is similar. Okimoto et al. (1992) suggested that A. suum and C. elegans may have shared a common ancestor as recently as 80 MYA. If this is true, the striking difference in AT-skew between these two taxa has arisen in a relatively short time. However, this divergence time is based on an estimated number of substitutions and an assumed rate of divergence derived from mammalian mtDNA studies. Because there have obviously been changes in the process of substitution along these lineages, as evidenced by the difference in AT-skew, this correction of sequence distance to divergence may be inadequate.

Insects

The D. yakuba (4.4 %GC) and A. mellifera (3.6 %GC) fourfold-degenerate sites are even more AT-rich than the nematode mtDNA. Drosophila mtDNA is the least skewed of all genomes considered here. Both GC-skew and AT-skew are small in this sequence and not significantly different from zero. The honeybee genome is more skewed than the fruitfly genome for both types of base pairs. Only one fourfold degenerate third codon position containing G was found in the Apis sample, compared to 14 sites containing C. These two insect taxa diverged from each other approximately 280 MYA (Crozier and Crozier 1993). Although both nematodes and insects have very AT-rich genomes, the skew patterns within and between these taxonomic groups are quite different.

Molluscs

This group is represented solely by the Mytilus sequence, which is 37.8% GC. Little AT-skew is observed in this genome; however, the GC-skew is of similar magnitude and direction to that found in nematodes. Thus far, this pattern is seen only among invertebrates, but is not a universally conserved feature.

Vertebrates

The %GC varies among vertebrates from 24.0% in lamprey, Petromyzon marinus, to 47.2% in human, Homo sapiens. Yet the negative GC-skew and positive AT-skew pattern is conserved in all these taxa. There is relatively little variation in the magnitude of either type of skew among vertebrate mitochondrial genomes, with the exception of a reduced AT-skew in lamprey mtDNA. GC-skew varies from approximately ~0.65 in the two teleost fish to approximately ~0.85 in the chicken. AT-
skews vary from 0.47 to 0.58 among vertebrates other than the lamprey (AT-skew = 0.16).

**Echinoderms**

Echinoderm genomes are less skewed overall than vertebrate mtDNAs. However, there is a substantial negative GC-skew in all four echinoderm taxa. *Arbacia lixula* mtDNA has less AT-skew than other echinoderm sequences and the direction of this skew differs from the common deuterostome pattern. This is primarily due to the increased %T in *Arbacia*. Observed asymmetries in the *Arbacia* substitution matrix (De Giorgi et al. 1991b) involve A → G transitions on the strand used in our analysis and do not explain the high %T.

**Correlations Among the Statistics**

Although the three statistics are logically independent, they are related by a common substitution matrix. Correlations might be expected to arise and may provide insight into the mechanisms generating variation in overall %GC and skew. At the very least, bivariate plots (Fig. 2) help to identify genomes with unique patterns of base composition. There is little correlation between %GC and either of the skew values (Fig. 2a,b), which is consistent with the idea that separate mechanisms control the overall composition of the double-stranded molecule and the strand-specific distribution of nucleotides (Thomas and Wilson, personal communication). We have already pointed out two instances where taxa with similar %GC have had dramatic differences in skew (nematodes and insects). Note also that the vertebrate genomes, while possessing a wide range of %GC, all have similar skews.

The relationship between GC-skew and AT-skew is shown in Fig. 2C. No sequence considered here exhibited highly positive values for both GC-skew and AT-skew or highly negative values for both skew measures. There is an interesting phylogenetic structure in the distribution of points in Fig. 2C. Vertebrates cluster in the upper left quadrant, echinoderms cluster around 0.0 AT-skew and -0.5 GC-skew, and other invertebrates are distributed in the region of positive GC-skew. The only exception is *Apis*, which clusters with the vertebrate taxa. *Ascocytis* is the only sequence to exhibit a strong negative AT-skew.

Earlier we mentioned that all these genes are believed to be on the same strand relative to the direction of replication, but that for several of the taxa, direct experimental evidence is not currently available. If the strands are reversed in nematodes and *Mytilus*, a skew pattern would be common to all animal mtDNAs except *Arbacia* and *Drosophila*. The relationship between AT-skew and GC-skew shown in Fig. 2C would then disappear, leaving all three statistics largely uncorrelated. Molecular studies to confirm the mechanisms of replication of additional mitochondrial genomes would be useful.

**Variation Within Genomes**

An important question is whether the values of these compositional statistics vary within mitochondrial ge-
nomes. As we subdivide the mitochondrial genomes to examine patterns of composition within a molecule, the number of appropriate sites for analyses of this kind becomes quite small, especially for less-frequent nucleotides. While some variation does exist among genes, a preliminary analysis indicates that intramolecular variation is small relative to the interspecific differences discussed in this paper. This is by no means intended to minimize the importance of intramolecular compositional patterns to understanding substitution mechanisms. Characterization of the overall patterns of composition is essential to the formation of testable hypotheses about the mechanistic origin of variation within genomes.

Predictions about the distribution of compositional variation both within and between genomes can be made from specific hypotheses about forces which shape the substitution process. For example, one hypothesis attributes strand-specific differences in the substitution matrix to differences in the damage spectra of single- and double-stranded DNA. This hypothesis was first put forward by Brown and Simpson (1982). In vitro, the rate of cytosine deamination is elevated approximately 200-fold in single-stranded DNA (Lindahl 1993). The asymmetric mechanism of replication (Clayton 1992) leaves some regions in a single-stranded state for as much as 30 min. The time spent single-stranded will vary in proportion to the distance from the replication origins (Thomas and Wilson, personal communication). This hypothesis has the potential to explain the common GC-skew pattern of vertebrates and echinoderms. Additional characterization of compositional patterns is necessary, however, to determine whether intramolecular compositional gradients consistent with this popular hypothesis exist.

Conclusions

It has long been recognized that the GC content of animal mitochondrial DNA varies widely among taxa and that the composition of the two strands is not equal in higher vertebrates, particularly mammals. It has not been widely appreciated that the two kinds of base pairs (GC and AT) can have separate behaviors. The statistics presented here will be useful for quantifying this variation in mitochondrial and other genomes. A clearer description of the patterns of nucleotide composition may ultimately lead to a better understanding of the biochemical mechanisms creating the patterns.

The biochemical mechanisms of mutation, repair, replication, transcription, and translation of mtDNA must all be taken into account as we search for the origin of compositional bias and skew in this molecule. Patterns of composition among individual genes and within fourfold degenerate codon families must be described in order to test a variety of hypotheses about composition and substitution, including the assumption that fourfold degenerate sites are not experiencing translational-level selection. We are currently investigating the use of log-linear modeling to examine the relationship between a number of characteristics of the mitochondrial genome and base composition.

Finally, there is a need for additional biochemical studies. Most of the information we have about mitochondrial replication, polymerase specificity, and DNA repair has come from studies of humans or mice. Studies of additional animal taxa are needed to provide comparative data so that we may better understand the evolution of animal mtDNA.

Acknowledgments. We thank W.K. Thomas for thoughtful discussions. This work was supported in part by National Science Foundation grants OCE-9116064 and BSR-9009795 to T.D.K.

References

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH,