A Comprehensive Analysis of Mammalian Mitochondrial Genome Base Composition and Improved Phylogenetic Methods

Andrew Gibson,* Vivek Gowri-Shankar,† Paul G. Higgs,* and Magnus Rattray†

*School of Biological Sciences, University of Manchester, Manchester, United Kingdom; †Department of Computer Science, University of Manchester, Manchester, United Kingdom; and ‡Department of Physics, McMaster University, Hamilton, Ontario, Canada

Phylogenetic analysis of mammalian species using mitochondrial protein genes has proved to be problematic in many previous studies. The high mutation rate of mitochondrial DNA and unusual base composition of several species has prompted us to conduct a detailed study of the composition of 69 mammalian mitochondrial genomes. Most major changes in base composition between lineages can be attributed to shifts between the proportions of C and T on the L-strand. These changes are significant at all codon positions and are shown to affect amino acid composition. Correlated changes in the base composition of the RNA loops and stems are also observed. Following up from previous studies, we investigate changes in the base composition of all 12 H-strand proteins and find that variability in proportions of C and T is correlated with location on the genome. Variation in base composition across genes and species is known to adversely affect the performance of phylogenetic inference methods. We have, therefore, developed a customized three-state general time-reversible DNA substitution model, implemented in the PHASE phylogenetic inference package, which lumps C and T into a composite pyrimidine state. We compare the phylogenetic tree obtained using the new three-state model with that obtained using a standard four-state model. Results using the three-state model are more congruent with recent studies using large sets of nuclear genes and help resolve some of the apparent conflicts between studies using nuclear and mitochondrial proteins.

Introduction

Mitochondrial genetics has become increasingly popular among evolutionary biologists since genome sequencing technology became efficient enough for rapid sequencing of mitochondrial genomes. Their small size, well defined set of genes, and increased mutation rate compared with nuclear genomes make them useful tools in phylogenetic analysis and a good alternative to nuclear benchmark genes. Yet, there is a danger that the rules that are generally applied to nuclear DNA do not apply to mitochondrial DNA (mtDNA) and mitochondrial genomes (mtGenomes) because of the significant differences in their nature.

Mammalian mtGenomes exist as closed circular strands and have a set of 13 protein-coding genes, two rRNA genes, and a full set of 22 tRNA genes. The two strands that make up the genome are most commonly known as the heavy strand (H-strand) and the light strand (L-strand) because of molecular weight differences arising from major differences in base composition between the two strands. Of the 13 protein-coding genes, 12 are on the H-strand and only one is on the L-strand. Noncoding regions are mainly limited to areas called the D-loop, thought to have functional roles in replication and transcription, and origin of replication of the L-strand (O_L), thought to have a functional role in replication (Shadel and Clayton 1997).

As new mammalian mtGenomes are sequenced, the focus of the resulting study is usually to produce an updated phylogenetic tree (e.g., Lin et al. 2002; Lin, Waddell, and Penny 2002; Reyes et al. 2004). Less often a study will analyze the base composition of a set of mtGenomes. Before so many mtGenome sequences were available, Perna and Kocher (1995) analyzed the nucleotide composition of 16 animal mtGenomes. Three measures were used on the third codon position of fourfold degenerate sites of the H-strand genes to understand underlying mutational patterns between the genomes. The first was simply the GC content of the sites, and the second and third were GC-skew and AT-skew calculated as (G − C)/(G + C) and (A − T)/(A + T), respectively. These statistics were designed as an indicator of differences between the two strands. A later study by Reyes et al. (1998) used the same statistics on mammalian mtGenomes to illustrate the strand heterogeneity between the two strands. This strand heterogeneity in base composition of mtGenomes presents the first major difference between a nuclear analysis and a mitochondrial analysis and means that, in mammals, only the 12 H-strand proteins are typically considered in any phylogenetic or base composition analysis. Schmitz, Ohme, and Zischler (2002) performed the most recent study of base composition in mammals that compares 26 mammalian mtGenomes. The focus of the study is on the sequencing of the Tarsius bancanus genome and the base composition among the Primates. They conclude that the higher Primates show a compositional shift from T to C and A to C, and these changes have an impact on the amino acid usage of the Primates as a result of a generally increased mutation rate that is restricted to anthropoids. Base composition in mitochondria has also been compared at the rRNA level. Springer and Douzery (1996) compare 49 complete mammalian rRNA genes, and findings include low variability of base composition across Mammalia, a higher percentage of adenine in loop regions than stem regions, and a high G+C composition in stem regions.

A second major difference between mitochondrial genomes and nuclear genomes is the replication mechanism. For a long time, it has been considered that mitochondrial genomes have an asymmetric replication mechanism (Clayton 1982; Shadel and Clayton 1997). During this process, the H-strand is replicated first, displacing the parental H-Strand and leaving it single stranded until replication

Key words: Mitochondria, phylogenetics, base composition, nucleotide bias, mammals, substitution model.

E-mail: magnus@cs.man.ac.uk.

Mod. Biol. Evol. 22(2):251–264, 2005
doi:10.1093/molbev/msi012
Advance Access publication October 13, 2004

Molecular Biology and Evolution vol. 22 no. 2 © Society for Molecular Biology and Evolution 2005; all rights reserved.
of the L-strand begins approximately two-thirds of the way around the genome. Several studies have correlated the relative amount of time that a gene on the H-strand is single stranded with base composition and have found a relationship between the two (Tanaka and Ozawa 1994; Reyes et al. 1998; Bielawski and Gold 2002; Faith and Pollock 2003). Interestingly, more recent studies have presented evidence of replication intermediates associated with more traditional bidirectional replication mechanisms (Holt et al. 2000; Yang et al. 2002; Bowmaner et al. 2003). Although this evidence is compelling, it does not provide an explanation for changes in base composition between genes that currently can be explained with the asymmetric replication mechanism. The resolution of which replication mechanism exists will surely aid our understanding of mutational patterns in mitochondria.

Strong changes in base composition lead to changes in amino acid composition (Foster, Jermiin, and Hickey 1997; Singer and Hickey 2000; Schmitz, Ohme, and Zischler 2002). If there are strong changes in base composition in mtGenomes, then standard amino acid and nucleotide-based phylogenetic models may be unable to compensate for these changes, particularly if they change at a variable rate between the species (Foster and Hickey 1999). In this study, we have performed a comprehensive study of changes in base composition and amino acid content across 69 mammalian mtGenomes. Because most mtDNA is coding, we have considered it in terms of five separate classes of functional site, three sites from the codon structure of protein-coding genes, and two from RNA-coding genes that represent nucleotides involved in either a helical or loop configuration. We expected the levels of selective pressure to be different for each of these classes of site and, therefore, have analyzed the base composition of each separately. Using the information we gathered from the base composition analysis, we have then measured the effect of a change in base composition on the amino acid composition.

Mitochondrial genes are used extensively in phylogenetic analyses and provide an important resource for phylogenetic analyses. Anatomical, paleontological, and molecular data now agree reasonably well at the level of mammalian orders, and recent studies in molecular phylogenetics also indicate that good progress is being made toward determining the supraregional relationship of mammals. However, there have been some striking inconsistencies along the way, particularly between studies using mtGenomes and those using mainly nuclear proteins. Certain species appear to be particularly troublesome in studies of mitochondrial proteins. The murid rodents and hedgehogs are often found at the base of the placental mammals, and the orders containing them are, therefore, often not found to be monophyletic using mitochondrial data sets (e.g., Arnason et al. 2002; Lin et al. 2002). However, studies using large data sets of predominantly nuclear genes (Madsen et al. 2001; Murphy et al. 2001a, 2001b; Deluc et al. 2002) do find support for monophyly of these orders, as do studies using the complete set of mitochondrial RNA genes (Hudelot et al. 2003). One factor that has been implicated as leading to spurious results is the differences in substitution processes and base composition across species. There have been some attempts to correct for this effect, for example by removing troublesome species or constraining certain well-established relationships (Lin et al. 2002) or increasing taxon sampling (Reyes et al. 2004). To correct for variation at the first codon position, Reyes et al. (2004) excluded Leucine synonymous sites in first codon positions after it was noticed that it varied considerably between species. Eliminating the compositional bias in this way improved the positioning of the troublesome species, although Eulipotyphla remain paraphyletic. Philips and Penny (2003) noted inconsistencies in the resolution of deep divergences in the mammalian tree obtained from mtGenomes and ascribed these to the large discrepancies in the relative frequency of T and C. They used RY-coding, in which purines and pyrimidines are lumped into composite states, to reduce these discrepancies and found improved resolution for the earliest branches of the tree (Phillips and Penny 2003). In this paper, we introduce a method that is similar in spirit to RY-coding but that retains information from transitions between purines and, therefore, improves resolution on shorter timescales.

Methods

Base Composition Analysis

Sixty-nine mammalian mtGenomes were chosen from the OGRe database (Jameson et al. 2003; latest version of OGRe available at http://ogre.mcmaster.ca). This set of species is the same set used by Hudelot et al. (2003), and the scientific names of species and NCBI accession numbers are given there. OGRe provides a facility to display codon usage statistics for each strand of each genome. The codon usage statistics for the 12 H-strand genes were processed in Microsoft Excel spreadsheets, where base composition was separated into first, second, and third codon positions.

For RNA base composition analyses, OGRe also provides alignments of the short-subunit and long-subunit rRNA genes that were created, with reference to the secondary structure. These alignments indicate whether an individual base is thought to be involved in a stem or a loop configuration in the resulting RNA molecule. The alignments were concatenated in the BioEdit sequence alignment program. In the concatenated alignment, there were 2,560 sites in total, of which 1,064 nucleotides were considered to be in helical regions (532 pairs). Gaps in the nonhelical regions were treated as missing data in the analysis. Base composition was ascertained using the base composition analysis function in BioEdit.

We used the PHASE phylogenetics package (described below) to produce maximum-likelihood base frequency estimates for all 12 H-strand protein genes in four separate groups of organisms. The four groups analyzed were Primates, Cetartiodactyla, Carnivora, and Lagomorpha/Rodentia. These four groups were selected primarily for their base compositional properties, as we shall see, the Primates and Lagomorpha/Rodentia have unusual changes in base composition between the species, whereas Carnivora and Cetartiodactyla represent sets of species that have less dramatic changes in base composition. For the purposes of this part of the study, the small sizes of the genes ND3 and ATP8 prompted us to combine these two genes with a neighboring gene: ND3 with ND4L.
and ATP6 with ATP6. The theoretical duration of the single-stranded state of the parental H-strand during strand asymmetric replication for each gene is as described by Reyes et al. (1998).

Phylogenetic Analysis

The nucleotide and amino acid sequence data of the 12 H-strand genes for all 69 genomes was obtained from the OGRE database. Amino acid sequences were manually aligned using the BioEdit software, followed by a nucleotide alignment based upon the amino acid alignment. Third codon positions were then stripped from the alignment to leave an alignment of 7,402 sites.

The phylogenetic analysis was carried out using the PHASE phylogenetic inference software for maximum-likelihood and Bayesian phylogenetic inference (available from www.bioinf.man.ac.uk/resources/phase). This package contains a number of substitution models for nucleotides and RNA base pairs. We used the Markov chain Monte Carlo (MCMC) Bayesian inference methods provided by the package. For this study, we have included a new substitution model in the package, which we describe below.

Substitution Models

The new model implemented is the most general time-reversible three-state model (GTR3) in which C and T are combined into a single composite pyrimidine state Y. This model is useful for describing DNA in which the C and T composition is highly variable across genes and across species. GTR3 is analogous to the widely used general time-reversible four-state model (GTR4). The GTR3 model has three independent exchangeability parameters $\alpha_{ik}$, determining the substitution rate between different states $i$ and $k$. To impose reversibility, the matrix of exchangeability parameters is symmetrical: $\alpha_{ik} = \alpha_{ki}$. The model has three frequency parameters $\pi_i$, and the instantaneous substitution rate between states $i$ and $k$ is defined to be $r_{ik} = \alpha_{ik} \pi_i$. Because the frequencies must sum to 1, the GTR3 model has five free parameters, which is four less than the standard GTR4 model.

Substitution rates and branch lengths cannot be evaluated independently, and substitution models are usually normalized so that the expected number of substitutions per site and per unit of branch length is one. This constraint effectively reduces the number of free parameters of a substitution model by one. When a three-state model is used, hidden transitions between the two pyrimidines C and T are excluded from this count. We have used different models for each of the first two codon positions. Standard tests (AIC and likelihood-ratio tests) show that there was overwhelming statistical support for using two separate models instead of a single model for both positions on this data set. We considered three different model combinations, GTR3-3, GTR4-4, GTR3-4, in which the first (second) number corresponds to the model used at the first (second) codon position. Separate models have also been used for each codon position in other studies (e.g., Delsuc, Phillips, and Penny 2003). It is not possible to use standard tests to compare these three models, because the likelihoods of the three-state and four-state model are not directly comparable. However, we were able to use a Cox test to test the hypothesis that the data were generated by GTR4-4. We used the consensus topology under GTR4-4 with maximum-likelihood branch lengths, and model parameters and alignments were simulated from this tree to compute the empirical sampling distribution of the log-likelihood ratio of GTR4-4 against the alternative models GTR3-4 and GTR3-3. When computing the likelihood ratio on the actual data, we removed any gapped sites to make the likelihood comparable to those of the simulated data. We found evidence to reject GTR4-4 in favor of GTR3-4 ($P = 0.02$) but no evidence for rejecting GTR4-4 in favor of GTR3-3. Unfortunately, it is not possible to carry out the Cox test with either of the other models as a null hypothesis, because the three-state model cannot be used to generate four-state data. However, we feel that the Cox test gives some support for using GTR3-4 over the other two models. We assume that the average substitution rate of the two models is related by a proportionality factor $c$, and we assume the same branch lengths for the tree under each model (Yang 1996). The model for the first codon position is normalized so that its average substitution rate is 1 and the second model is normalized so that its average substitution rate is equal to $c$.

We model rate heterogeneity using the discrete gamma model of Yang (1994). This model assumes that the average substitution rate across different sites is distributed according to a gamma distribution. The mean of this distribution is equal to the average substitution rate of the substitution model, and its shape is controlled by a single parameter. To make the method tractable, we approximate this distribution by six equiprobable rate categories. We use two gamma distributions, with independent shape parameters for each codon position.

Bayesian Phylogenetic Inference

Bayesian inference methods are becoming increasingly popular in molecular phylogenetics. Recent reviews of Bayesian phylogenetics and MCMC techniques are provided by Huelsenbeck et al. (2001) and Holder and Lewis (2003). These methods have clear computational advantages over standard statistical methods, and they permit the exploration of a complex model space in a reasonable amount of processing time. MCMC algorithms implemented in PHASE have previously been described by Jow et al. (2002), and they have been used with minor modifications here (details are provided in the PHASE manual, available from www.bioinf.man.ac.uk/resources/phase). For each MCMC cycle, an attempt is made to change either (1) the topology of the tree using a local nearest-neighbor interchange or a subtree pruning and regrafting proposal, (2) one parameter of the substitution model, or (3) one branch length. In this paper, we used different proposal probabilities for these moves to tailor the mixing of the variable number of free parameters in the substitution model. Truncated uniform priors are used for branch lengths and for all substitution model parameters except frequencies. To propose new values for these parameters, we use Gaussian distributions centered at the current value, with reflecting boundaries at zero and
the prior upper bound. We assume a flat Dirichlet prior for frequencies, and we use a Dirichlet proposal distribution centered at the current vector of frequencies. Parameters determining the variance of these proposal distributions are adjusted to control the mixing behavior of the chain. In practice, the burn-in period is used to tune these parameters and to have a reasonable acceptance rate (20% to 25%) during the sampling.

For each run, the burn-in period was 750,000 cycles, and this was found to always be clearly sufficient for the likelihood and the substitution model(s) parameters to reach equilibrium. After the burn-in, 15,000 trees were sampled every 100 cycles during the sampling period (1,500,000 cycles). For each experiment described in this paper, four separate MCMC runs were performed. Results of the four independent chains (with four independent initial configurations) were compared to give us a high level of confidence that the equilibrium was reached. Mixing behavior of parameters were compared between chains and clade supports (i.e., the Bayesian posterior probabilities supporting each clade) were found to be very similar between the four runs in all our experiments. The 60,000 samples obtained from the four chains were then concatenated to produce the consensus results shown here.

Results and Discussion

Base Composition Analysis

We have analyzed the base composition of five classes of functional site over 69 mammalian mtGenomes. These classes are the three codon positions in the 12 H-strand protein-coding genes and the loop regions and helical regions in the short-subunit and long-subunit rRNAs.

Protein Genes—Third Codon Position

Figure 1 shows the L-strand DNA base composition of the third codon position for the 69 genomes. By ordering the species by increasing percentage of C, a very large variation in composition from 18% to 45% is highlighted. Also, as the percentage of C increases across the species, a significant corresponding decrease in the percentage of T can be observed ($r = -0.8683$, $P = 1.0E-22$). The percentage of G at the third codon position on the L-strand is consistently low (less than 10%), and the percentage of A is always high (more than 35%). As the percentage of C rises above 40%, there also appears to be a corresponding decrease in the percentage of A, which may account for a significant correlation between the proportions of these two bases ($r = -0.4958$, $P = 6.4E-06$). There are no other significant correlations between any other pair of bases at this position.

The strong changes in composition involving C and T are large and significant, prompting us to look at other classes of site within the genome for similar changes. The relationship between C and A is not as well defined and does not seem as strong as the relationship between C and T. It is interesting, however, that the relationship only becomes apparent when the proportion of C reaches a certain level and T composition is at about 20%. This indicates that under these conditions, C can be substituted for A as well as T, although when T is more abundant, the relationship between C and T is much more pronounced.

Protein Genes—First and Second Codon Positions

Figures 2A and B show the base composition of the first and second codon positions for the 69 genomes, sorted by increasing levels of C. We see similar changes in the compositions of C and T as in the third codon position. Although there was no reason to believe that the compositional bias seen in the third codon position was anything more than a neutral compositional drift, the changes we see here indicate that the source of the change affecting C and T in the third codon position is also affecting first and second codon position composition, and hence amino acid composition.

The composition of C and T at the first codon position does not vary as much as at the third codon position, but the correlation between them is more significant ($r = -0.9326$, $P = 4.17E-32$). Similarly, compositions of C and T at the second codon position vary even less but have an even more significant relationship ($r = -0.9535$, $P = 1.90E-37$). This does not imply that the relationship between C and T is weaker at the third codon position compared with the first or second codon position, but illustrates that increased variability arising from neutral substitutions in the third codon position adds noise to the data. Table 1 shows the range of percentage composition for each base at each codon position. If we take these values as a measure of the amount of purifying selection acting at each site, then it would appear that the second codon position is under greater purifying selection than is the first codon position. It is expected that selection will be greatly reduced at the third codon position because of the
redundancy in the code. Less expected are the differences in the apparent amount of purifying selection between the first and second codon positions. The composition of the second codon position is incredibly rigid, with very little deviation in composition across the species, as can be seen in figure 2B. Despite this, C and T still vary together by a noticeable amount, highlighting how strong the pressure for changes between these two bases appears to be. In comparison, the composition of the first codon position is much less rigid. Some of the variation is expected, as Leucine has two codon families, CTN and TTR, giving the first codon position some redundancy in C and T transitions. This redundancy does not account for the increased variability of all bases compared with the second codon position. We will see later that the substitution rate is indeed much lower on average at the second codon position.

**rRNA Genes—Helices and Loops**

Figures 2C and D show the base composition of the helical and loop regions of the rRNA genes, respectively. Once again, the compositions of C and T fluctuate across the species in the same way as in the three codon positions but with minor differences. This observation seems to be in contradiction to the observation made by Springer and Douzery (1996) that base composition is fairly uniform across Mammalia, as there are significant variations in C and T.

Selection is expected to be relatively strong in the rRNA helical regions to maintain RNA structure and function, and it is, therefore, interesting that we find the base composition varies in the same way as other functional sites. Even though the change is not on the same scale (around a 4% variation across species), it is
very significant \( (r = -0.9559, P = 3.36\times10^{-38}) \). rRNA helical sequences have a base pairing pattern needed to form the rRNA secondary structure, meaning that changes in C and T composition also bring about changes in the A and G composition. In figure 2C, we see that there is a large fraction of G in the helical regions and that G is the most frequent base in about half the species. This is in agreement with observations made by Springer and Douzery (1996) and is in marked contrast to both the rRNA loops and the protein-coding sequences, where G is the least frequent base. We also see that the fraction of G in the helices increases as the fraction of C increases, whereas this trend is not seen in either the loops or the protein sequences. This indicates selection for maintaining a stable RNA secondary structure. GC pairs are the most thermodynamically stable, and G also has the ability to pair with U, which explains why G has a significantly higher frequency than C in the helices. There seems to be a mutational bias away from G (as evidenced by the very low G content in the third positions). However, the trends seen in figure 2C show that there is strong selective pressure to maintain secondary structure. The amount of variability in the changes of the percentages of A and G can be attributed to certain mismatched base pairs that are allowed in rRNA helices. Selection for stabilizing RNA structure is seen in many types of structural RNA sequences (Higgs 2000).

rRNA loops may be expected to have weaker selective pressure acting on their composition. The correlations between levels of C and T in the rRNA loop regions are not quite as significant as in the third codon position, but are still highly significant \( (r = -0.8161, P = 3.36\times10^{-18}) \). There is also a correlation between C and A in the loop regions \( (r = -0.6561, P = 3.51\times10^{-10}) \) that is slightly more significant than at the third codon position. An unusually high percentage of A is also apparent in rRNA loop regions, consistent with previous studies that suggest this is necessary for hydrophobic interactions with proteins (Gutell et al. 1985; Springer and Douzery 1996).

**Congruent Changes at All Sites**

Thus far, we have considered each category of site independently, but we have not considered whether changes are correlated across site categories. Figure 3 shows plots of third codon position percentage of C against the percentage of C in each of the other four categories of site that we have studied. Table 2 shows the regression line formulas and significance values for each correlation. It is clear from figure 3 that all of the five categories of site are being affected by the same compositional change in the same way, and the magnitude of the change is limited by the amount of purifying selection acting on each site based on its function.

We can now say that the primary change that occurs between species involves changes in the percentages of C, and, hence, T, on the L-strand that appears to arise from a mutational bias that varies from organism to organism. The evidence provided so far is a clear indicator that the changes in the composition involving C and T between the organisms are genome-wide, as selection pressures acting upon protein-coding sequences are expected to be independent of selection pressures acting upon rRNA sequences.

**Base Composition Affects Amino Acid Composition**

Figure 4 is a summary of how the changes in base composition across the mtGenomes affect the amino acid composition. Figure 4A shows the vertebrate mitochondrial translation table. An amino acid has a gray background if its frequency significantly correlates with the frequency of C at the third codon position across the species and a number that is the correlation coefficient for that relationship. Arrows have been added to the table to

<table>
<thead>
<tr>
<th>Codon Position</th>
<th>%A Range</th>
<th>%C Range</th>
<th>%G Range</th>
<th>%T Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>3.97</td>
<td>8.59</td>
<td>3.72</td>
<td>7.28</td>
</tr>
<tr>
<td>Second</td>
<td>1.58</td>
<td>4.72</td>
<td>1.22</td>
<td>3.62</td>
</tr>
<tr>
<td>Third</td>
<td>14.07</td>
<td>28.05</td>
<td>7.38</td>
<td>23.53</td>
</tr>
</tbody>
</table>

**Table 2**

**Regression Line Formulas, Correlation Coefficients, and \( P \) Values of Lines Plotted in Figure 3**

<table>
<thead>
<tr>
<th>mtDNA</th>
<th>Regression Line</th>
<th>( r )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>First codon position: ( y = 0.2561x + 17.221 )</td>
<td>0.9188</td>
<td>1.84E−29</td>
<td></td>
</tr>
<tr>
<td>Second codon position: ( y = 0.0997x + 23.681 )</td>
<td>0.7713</td>
<td>2.79E−15</td>
<td></td>
</tr>
<tr>
<td>rRNA loop: ( y = 0.4052x + 7.9867 )</td>
<td>0.9113</td>
<td>3.27E−28</td>
<td></td>
</tr>
<tr>
<td>rRNA helix: ( y = 0.1034x + 20.535 )</td>
<td>0.6662</td>
<td>1.55E−10</td>
<td></td>
</tr>
</tbody>
</table>
indicate the inferred change in amino acid as $T$ changes to $C$. It is known that leucine codon usage varies across the species (Reyes et al. 2004), but from figures 4A and B, it is now obvious that that change is strongly linked to the overall changes in base composition seen in this study. Not so well characterized are changes in the compositions of other amino acids across the species.

It is quite striking that the relationship between base composition and amino acid composition is so strong, and the levels of more amino acids than just leucine are significantly affected by this relationship. Figure 4B illustrates the relationship between the proportion of $C$ at the third codon position and the amount of CTN leucine ($r = 0.910, P = 6.06E–28$). Figure 4C shows the correlation between the percentage of proline and the amount of $C$ at the third codon position ($r = 0.737, P = 1.73E–13$). Although the total percentage of proline only varies by 1.4% of all amino acids, proline only represents 5.5% of all amino acids, making this a relatively large variation. Figure 4D shows the average percentage of each amino acid, which, in terms of figure 4A, shows how the most significant changes in composition are made by the most-represented amino acids. Leucine is by far the most-represented amino acid, and it also varies the most in response to changes in base composition. Curiously the levels of $V$ and $A$ do not significantly change with base composition, which seems to suggest that the strength of selection acting to prevent $V/A$ replacements is greater than the strength of selection acting on other amino acid pairs such as $I/T$ and $L/P$.

**Trends Across Species and Within Genomes Species**

So far we have shown that the composition of $C$ and $T$ differs between the organisms, but it is still not apparent...
what causes these differences or whether there is a common factor linking species with similar base composition properties. Figure 5 is a phylogenetic tree adapted from Hudelot et al. (2003) of all 69 species studied. The tree was created using the complete set of mitochondrial tRNA and rRNA genes. We have added a specific color to each organism representing the percentage of C in the third codon position to illustrate the distribution of compositional bias across the species.

Among the species that stand out as having a high percentage of C are a group of eight Primates that all have between 39% and 45% of C at the third codon position. This observation is consistent with previous observations made about the base composition of higher Primates (Schmitz, Ohme, and Zischler 2002). At the opposite end of the scale are the Western European hedgehog and the moon rat, which have approximately 18% C at the third codon position. The overall distribution of color across the rest of the species seems susceptible to change and does not appear to follow any pattern, except for closely related species that often share very similar compositions. Significant changes, when they do occur, happen in an unpredictable way (i.e., at a variable rate) and over quite short timescales. Two prime examples of this are the murid rodents (Chinese vole, Norway rat, and house mouse) and the lagomorphs (rabbit, European hare, and pika). Each member of these two groups has a substantially different proportion of C at the third codon position (37%, 33%, and 28% in murid rodents and 31%, 27%, and 41% in lagomorphs). This lack of clear correlation between compositional bias and phylogenetic position is quite striking but does not give any obvious indication as to what factor(s) may be responsible for such variations.
groups. This is because of the nature of the overall base composition changes not only at a different rate between species but also at a different rate between genes, which leads to inconsistency of assumptions made by phylogenetic methods.

Phylogenetic Inference

Recent studies in molecular phylogenetics indicate that good progress is being made toward determining the supraordinal relationship of the placental mammals. However, there have been some striking inconsistencies along the way, particularly between studies using mtGenomes and those using mainly nuclear proteins. One important feature emerging from studies of nuclear proteins is that there is growing support for four major supraordinal clades, often referred to as Afrotheria, Xenarthra, Supraprimates (or Euarchontoglires), and Laurasiatheria (Madsen et al. 2001; Murphy et al. 2001a, 2001b). There is also strong support for the sister relationship of the latter two clades, although the overall relationship of the four clades and their root position remains unresolved (Delsuc et al. 2002). This emerging picture of the supraordinal placental phylogeny has been contradicted by studies of large sets of mtGenomes (e.g., Arnason et al. 2002) and unconstrained trees with outgroup species included (Lin, Waddell, and Penny 2002). Some of these inconsistencies have now been resolved by improved taxon sampling (Reyes et al. 2004), removing or constraining troublesome species (Lin, Waddell, and Penny 2002), focusing on RNA-encoding genes (Hudelot et al. 2003), or removing particular sites in the alignment to reduce compositional differences (Reyes et al. 2004). However, it is important to improve phylogenetic methods so that they are much more consistent across different data sets, as this will help improve confidence in the results produced by these methods when little prior information is available.

One possible reason for the inconsistencies between studies using predominantly nuclear genes and those using mitochondrial genomes is that differences in nucleotide composition across genes and species are biasing the results. Such differences in composition are inconsistent with the assumptions made by most maximum-likelihood and Bayesian phylogenetic inference methods. We have shown here that C and T composition varies significantly at functional sites in coding DNA, both across the genome and between the genomes of different species. This variation in composition is significant in the first two codon positions, although it is more marked in the first codon position, and results in significant variation in both codon and amino acid composition. We can, therefore, expect standard phylogenetic methods working at the level of codons, proteins, or nucleotides to be adversely affected.

To avoid the problem of variation in C/T composition, we have implemented a three-state substitution model, GTR3, in the PHASE phylogenetics package (see Methods for details). This is a three-state model in which T

![FIG. 6.—Maximum-likelihood base frequency estimates of C (A) and T (B) at the third codon position in four groups of species (1 = Primates, 2 = Cetartiodactyla, 3 = Carnivora, 4 = Lagomorpha/Rodentia), produced using the mlphase program of the PHASE software package. Genes in each group are in order of increasing theoretical single-stranded time during a strand asymmetric replication mechanism as defined in Reyes et al. (1998). From left to right: COX1, COX2, ATP6/ATP8, COX3, ND3/ND4L, ND4, ND1, ND5, ND2, and CYTB.](image_url)
and C are combined within a single pyrimidine state Y. The model is similar in spirit to the two-state RY-coding model in which only transversions are considered, with purines and pyrimidines lumped into composite states. RY-coding was recently used to overcome bias caused by compositional differences in mammalian mitochondrial sequences, and it was found to effectively resolve some of the earliest branches of the mammalian tree (Phillips and Penny 2003). However, we are also interested in other parts of the tree, and much useful information is lost by ignoring transitions altogether. Using a two-state model in place of our three-state model always results in significantly poorer resolution for all but the earliest branches in the tree. An alternative strategy for reducing bias caused by compositional effects is to exclude the sites from the alignment showing the greatest variability. This method was used by Reyes et al. (2004), who removed the first codon position from leucine-synonymous sites in their alignment. Because leucine is the most common amino acid, this resulted in them having to remove 28% of first codon positions from their alignment. We have shown above that it is not only leucine that is affected by compositional variation. Our method does not require the removal of any columns of the alignment and also compensates for changes affecting bases coding for the other amino acids.

We use different models for the first and second codon positions and we have carried out a separate analysis for three combinations, which we denote GTR3-3, GTR3-4, and GTR4-4, where the first (second) number corresponds to the model used in the first (second) codon position. In figure 7 we show the consensus tree for the analysis of 69 mammalian mtGenomes using the GTR3-4 model combination, which was the model favored by a Cox test (see Methods for details). Consensus trees for GTR3-3 and GTR4-4 are given in the Supplementary Material online. Numbers show the percentage of posterior probability for each clade; that is, the percentage of samples in the MCMC output containing that clade. Nodes without numbers are supported with 100% of posterior probability. Results obtained using GTR3-3 and GTR3-4 are quite similar, and with both models we find strong support for the same four main supraordinal clades found by Madsen et al. (2001) and Murphy et al. (2001a, 2001b). We also find highest posterior probability for the sister relationship of Supraptimates and Laurasiatheria (76% for GTR3-4 as shown in figure 7 and 86% with GTR3-3), in agreement with these studies and with Delsuc et al. (2002), although the level of support is not as high as our previous study using all of the RNA genes from the same set of species (Hudelot et al. 2003). The overall grouping of the four main clades is not well resolved by either model, although highest posterior probability places the root of the placentals on the branch to Afrotheria, again in agreement with Madsen et al. (2001), Murphy et al. (2001a, 2001b), and Delsuc et al. (2002). Our results for the early branches of the mammalian tree are, therefore, congruent with these studies using both the GTR3-3 and GTR3-4 combination of models. Under both models, our posterior probability support for the four main clades is higher than that obtained by Reyes et al. (2004) using a larger set of species, and they found highest posterior probability for an arrangement with Xenarthra as sister to Supraptimates, in contrast to the results of Madsen et al. (2001), Murphy et al. (2001a, 2001b), and Delsuc et al. (2002).

Resolution is quite good at the level of orders, with most of the established mammalian orders found to be monophyletic. However, there are some differences between the results obtained using GTR3-3 and GTR3-4. Within Supraptimates we find lagomorphs and rodents to be monophyletic sister clades using GTR3-4, whereas these orders mix together using GTR3-3, and the murid rodents form a clade with the tree shrew, a member of Scandentia. All models find the same arrangement for the primates and the Malayan flying lemur, with the flying lemur separating the anthropoids from the other primates. A similar sister relationship between the anthropoids and flying lemur is found by Reyes et al. (2004), which is perhaps not surprising, because we find that the result is robust to the particular choice of substitution model used.

Within Laurasiatheria the relative positioning of the orders is not very well resolved. The position of the long-tailed pangolin is quite variable under GTR3-4, with highest posterior support for positioning at the base of the carnivores, in agreement with Reyes et al. (2004), while with GTR3-3, it is found to be a sister to Chiroptera. The position of the pangolin was not well resolved in Madsen et al. (2001) or Murphy et al. (2001a), but with a larger data set, positioning at the base of the carnivores is favored. Murphy et al. (2001b) and Amrine-Madsen et al. (2003) find a large deletion providing compelling evidence for this arrangement. Eulipotyphla are found to be paraphyletic under GTR3-4 (see figure 7) with the hedgehog and moon rat basal within Laurasiatheria, whereas under GTR3-3, they are found to be monophyletic with a high posterior probability.

Results using the standard four-state GTR model for both codon positions (GTR4-4 [see Supplementary Material online]) are far less congruent with the studies using predominantly nuclear genes or those using mtRNA genes (Jow et al. 2002; Hudelot et al. 2003). Laurasiatheria are no longer found to be monophyletic, and the hedgehogs are found at the base of the placental mammals. This is similar to the results found by others using mitochondrial proteins (Arnason et al. 2002), and these species are known to be problematic in studies using mitochondrial proteins. There is weakened support (74%) for monophyly of the Supraptimates, as the position of the murid rodents is variable. Afrotheria and Xenarthra are found to be monophyletic with high posterior probability, but their relative positioning is different from that found using GTR3-3 and GTR3-4. Supraptimates are separated from Laurasiatheria, and there is strongest support for them forming a clade with Xenarthra and Afrotheria. Surprisingly, we find some support for monophyly of the rodents...
whereas previous studies of mitochondrial proteins using standard protein and nucleotide substitution models have often found highest support for a paraphyletic arrangement (Arnason et al. 2002; Lin, Waddell, and Penny 2002), with the murid rodents often found at the base of the placentals. It appears that by using different models for the first and second codon positions, we increase the support for monophyly of the rodents and reduce the tendency for the murid rodents to move towards the root of the placentals. When we use a single GTR4 model for both codon positions, we find a similar tree to Arnason et al. (2002), with strong support for the murid rodents being basal to the placentals and the rodents being paraphyletic. There is overwhelming statistical support for using two separate models, and this may be explained by the large difference in composition between codon positions (see figures 2A and B) and by the large difference in substitution rate between codon positions. Using both the GTR3-3 and GTR4-4 models, we find that the second codon position evolves at about 0.3 times the rate of the first codon position (mean posterior estimate for parameter c was 0.28 for both models). This confirms our earlier observation that there appears to be relatively strong purifying selection on average at the second codon position when compared with the first codon position. The increased substitution rate at the first codon position is not caused by the increased redundancy from the alternative classes of leucine codon, because the GTR3-3 model removes this redundancy, and the relative rates are almost identical for both model combinations.
Other methods have been proposed for reducing errors caused by variations in composition. One such method is the LogDet transform (Lake 1994; Lockhart et al. 1994) and it is, therefore, interesting to observe how trees constructed using LogDet distances perform on the present data set. We used the implementation in PHYLIP (Felsenstein 1989) and created a neighbor-joining tree with bootstrap support for each clade. The resulting tree is quite similar to that obtained using the GTR4-4 model, and there is very weak support for most of the early branches. The hedgehogs are found at the base of the placentals, whereas the positioning of Xenarthra is highly variable, with greatest support placing it as a sister to Laurasiatheria. Monophyly of the Supraprimates is supported by only 18% of bootstrap trees. Thus, the LogDet method does not appear to deal with the problems of the current data set very well. An alternative approach for dealing with compositional differences is to explicitly model them using a nonstationary substitution model. For example, Galtier and Gouy (1998) modeled the variation in GC-composition of RNA genes using a simple nonstationary model in which every branch of the tree was associated with a parameter determining the corresponding equilibrium GC-composition for that branch. An advantage of using nonstationary models is that one can model more general changes in the substitution process and not just compositional changes. However, we note the composition also varies across genes in the present study, and such variations are not accounted for by a nonstationary model. We have recently implemented some nonstationary substitution models in the prototype version of PHASE, and it will be interesting to investigate whether such models are generally useful.

Bayesian methods have been criticized on the basis that they can lead to overconfidence in an incorrect hypotheses under certain circumstances, and it has been shown that bootstrap support levels and Bayesian confidence intervals, although highly correlated, may be related in a variable and nonlinear way (Buckley 2002; Waddell, Kishino, and Ota 2002; Suzuki, Glazko, and Nei 2002; Alfaro, Zoller, and Lutzoni 2003; Douady et al. 2003; Taylor and Piel 2004). Bayesian methods give a better measure of confidence in situations in which the substitution model is a good fit to the data (Wilcox et al. 2002), and Bayesian methods also provide the most efficient use of data in this case (Alfaro, Zoller, and Lutzoni 2003). However, in practice, it is likely that the substitution model used is highly idealized and far from the true evolutionary process, in which case posterior probabilities are often higher than bootstrap support values and may overestimate confidence (Douady et al. 2003; Taylor and Piel 2004). It was noted by Taylor and Piel (2004) that very strongly supported nodes (those with greater than 99% of posterior probability) result in very few false-positive predictions and we find greater than 99% support for the four major supraordinal clades, suggesting that these clades would indeed also be strongly supported by a maximum-likelihood analysis. The current version of the PHASE software includes maximum-likelihood methods but does not support fast topology-search methods for use in a maximum-likelihood analysis. It would, therefore, be useful to include fast distance-based and maximum-likelihood methods within PHASE, so that we can provide bootstrap supports for maximum-likelihood analysis, as well as posterior probabilities. More theoretical work is also required on this issue because there is a real danger that the acknowledged advantages provided by the Bayesian MCMC approach, in terms of greatly improved flexibility and computational efficiency (Holder and Lewis 2003), could be overshadowed by this issue.

Conclusions

Mammalian mitochondria exhibit striking variations in base composition both within genomes and across species. We have carried out a comprehensive analysis of base composition in 69 mammalian mtGenomes. Variations can mainly be attributed to large changes in L-strand C and T. This variation can be observed at all three codon positions and in the tRNA loops and helices, indicating a genome-wide correlated trend. We have implemented a new three-state substitution model that reduces the bias in phylogenetic inference because of these large variations. We also find significant differences in composition and substitution rate at the first and second codon positions, and the use of separate models for each site significantly improves the fit of the data. Phylogenetic trees obtained using the new model show increased congruence with recent studies of large sets of nuclear genes, suggesting that many of the apparent conflicts between results obtained using nuclear and mitochondrial proteins may be a consequence of the lack of compensation for variation in base composition in mtGenomes. GTR3 attempts to remove effects caused by variation in T and C from the phylogenetic reconstruction and, therefore, artificial relationships between taxa. This model deals with the bias in a consistent fashion at all sites and for an unselective set of species, rather than resorting to the removal of specific sites or species. As with RY-coding, we obtain improved resolution for the early branches of the tree, but by retaining purine transitions in the model, resolution at shorter timescales is much improved. The apparent success of this method illustrates how applying a generic model to a specific system is not always the best approach, especially in the case of mammalian mitochondria, where it has been assumed that better trees will come from better taxon sampling. In fact, the introduction of species with extreme nucleotide bias could easily produce unusual phylogenetic inferences using the traditional method of tree construction, as with the addition of Eulipotyphla in Lin et al. (2002).

It is now obvious that the proportions of C and T both change between different species and that changes are congruent at all sites examined within the genome, consistent with a directional mutation pressure and in agreement with similar analyses conducted by Schmitz, Ohme and Zischler (2002). At the third codon position and rRNA loop categories of site, when C is elevated to a particularly high level and T is depleted, A appears to be substituted for C instead of T, producing significant correlations between C and A at these sites. A speculative explanation for these observations would be a directional
mutational process related to a strand-asymmetric replication mechanism or transcription, as we have shown the proportions of C and T at the third codon position in each gene change in relation to their position on the genome.

Other possible explanations might include differences in the efficiency or nature of mtDNA repair mechanisms between organisms (see Bohr [2002] for a review) or survival of the mRNA in the oxidative conditions of the mitochondria. The observed variations in C and T may reflect a similar type of selection pressure previously suggested for low %G (Reyes et al. 1998) acting on the base composition of mRNAs to preserve their life span, rather than preservation of the DNA. This could be caused by variations in the oxidative conditions in the mitochondria between species. Another target for study is the gene ND6, which is the only gene on the L-strand. This gene is largely ignored by most studies because it inherits the strand bias of the L-strand and, therefore, is not compatible with studies of the H-strand genes. We applied similar tests to ND6 in the hope that we would be able to identify a significant difference in its composition that might arise from it being located on the opposite strand to the rest of the protein-coding genes, but because of its small size and, hence, lack of data, these results were inconclusive.

Acknowledgments

A.G. gratefully acknowledges a BBSRC studentship award. We would like to thank Mark Springer and two anonymous reviewers for making a number of useful comments on a preliminary version of the paper and also thank Steve Oliver for useful discussions.

Literature Cited


Mark Springer, Associate Editor

Accepted October 4, 2004