Compositional differences within and between eukaryotic genomes

SAMUEL KARLIN AND JAN MRÁZEK

Department of Mathematics, Stanford University, Stanford, CA 94305-2125

Contributed by Samuel Karlin, July 9, 1997

ABSTRACT Eukaryotic genome similarity relationships are inferred using sequence information derived from large aggregates of genomic sequences. Comparisons within and between species sample sequences are based on the profile of dinucleotide relative abundance values (The profile is \( \rho_{XY} = f_{XY}/f_Y \) for all \( XY \), where \( f_X \) denotes the frequency of the dinucleotide \( X \) and \( f_Y \) denotes the frequency of the dinucleotide \( XY \), both computed from the sequence concatenated with its inverted complement). Previous studies with respect to prokaryotes and this study document that profiles of different DNA sequence samples (sample size \( \geq 50 \) kb) from the same organism are generally much more similar to each other than they are to profiles from other organisms, and that closely related organisms generally have more similar profiles than do distantly related organisms. On this basis we refer to the collection \( \{ \rho_{XY} \} \) as the genome signature. This paper identifies \( \rho_{XY} \) extremes and compares genome signature differences for a diverse range of eukaryotic species. Interpretations on the mechanisms maintaining these profile differences center on genome-wide replication, repair, DNA structures, and context-dependent mutational biases. It is also observed that mitochondrial genome signature differences between species parallel the corresponding nuclear genome signature differences despite large differences between corresponding mitochondrial and nuclear signatures. The genome signature differences also have implications for contrasts between rodents and other mammals, and between monocot and dicot plants, as well as providing evidence for similarities among fungi and the diversity of protists.

Local and global compositional heterogeneity is recognized on many scales in eukaryotic genomes, including variation in G+C content (e.g., isochore compartments, coding vs. non-coding), mobile insertion elements, characteristic centromeric satellite and telomeric repeats, CpG (=CG) suppression in vertebrates, and methylation patterns. Our recent studies of genomic sequence data have demonstrated that (i) the dinucleotide relative abundance values (defined below) of different sequence samples of DNA from the same organism are generally much more similar to each other than they are to sequence samples from different organisms and (ii) related organisms generally have more similar dinucleotide relative abundance values than do distantly related organisms (1). Dinucleotide relative abundance values are equivalent to the “general designs” derived from biochemical nearest-neighbor frequency analysis (2, 3). These highly stable DNA doublet forms suggest that there may be genome-wide factors, such as functions of the replication and repair machinery, context-dependent mutation rates, DNA modifications, and base-step conformational tendencies that impose limits on the compositional and structural patterns of a genomic sequence. The set of dinucleotide relative abundance values constitutes a “genomic signature” (1, 4) that may reflect the influence of such factors.

Relative Abundance Values

A standard assessment of dinucleotide bias is through the odds ratio \( \rho_{XY} = f_{XY}/f_Y \), where \( f_X \) denotes the frequency of the dinucleotide \( X \) and \( f_Y \) denotes the frequency of the dinucleotide \( XY \). The formula for \( \rho_{XY} \) is modified for double-stranded DNA by calculating the odds ratio \( \rho_{XY} \) for the given DNA sequence concatenated with its inverted complementary sequence. The set of \( \rho_{XY} \) values is referred to as the dinucleotide relative abundance profile.

A measure of difference between two sequences \( f \) and \( g \) (from different organisms or from different regions of the same genome) is the average absolute dinucleotide relative abundance difference calculated as \( \delta^*(f,g) = 1/16 \Sigma_{XY}|(\rho_{XY}(f) - \rho_{XY}(g))| \), where the sum extends over all dinucleotides.

DNA Sequence Samples

Large collections of genomic DNA sequences were extracted from GenBank (Release 101.0, June 1997) for 32 diverse eukaryotic organisms (see legend to Table 1). We restricted attention to species for which at least 100 kb of nonredundant sequence data was available. Most sequence aggregates include several contigs \( \geq 10 \) kb and often \( \geq 30 \) kb. Apart from complete chromosomes, the sequences in each species ensemble were culled of duplications. It appears that simple eukaryotes like yeast, nematodes, and Arabidopsis consist mostly of closely juxtaposed gene sequences interrupted with rare mobile DNA elements. The current human collection (all contigs \( \geq 80 \) kb) tends to be biased toward genes of medical interest. The D. melanogaster data set is replete with developmental genes. The data collections are certainly unrepresentative of the scope of invertebrate, plant, and protist taxa. For example, the dicots, apart from A. thaliana, cover only a few species and the monocots are restricted to a few grasses.

Dinucleotide Compositional Extremes

From statistical theory and data experience, a dinucleotide relative abundance may be conservatively described as significantly low if \( \rho_{XY} < 0.78 \) and significantly high if \( \rho_{XY} \geq 1.23 \). We distinguish extremes of dinucleotide relative abundances as follows: extremely high, symbolically +++, \( \rho_{XY} \geq 1.50 \); very high, ++, \( 1.30 \leq \rho_{XY} < 1.50 \); significantly high, +, \( 1.23 \leq \rho_{XY} < 1.30 \); marginally high, (+), \( 1.20 \leq \rho_{XY} < 1.23 \); extremely low, --, \( \rho_{XY} \leq 0.50 \); very low, --, \( 0.50 < \rho_{XY} \leq 0.70 \); significantly low, -, \( 0.70 < \rho_{XY} \leq 0.78 \); marginally low, (–), \( 0.78 < \rho_{XY} \leq 0.81 \).

The following trends were observed (see Table 1). (i) TA is broadly underrepresented in eukaryotes (and prokaryotes) generally in the range \( \rho_{TA} \approx 0.61–0.81 \). The
Each sequence collection was cleaned by removing duplicate copies of the same gene as identified in the sequence annotations. Species names are abbreviated as follows: Vertebrates: homsa (Homo sapiens, human), aggregate sequence length 3,326 kb, minimum contig size 80 kb; bost (Bos taurus, bovine, 136, 10); sucss (Sus scrofa, pig, 142, 10); rorc (Rattus norvegicus, rat, 656, 10); and musm (Mus musculus, mouse, 1,065, 50); ratio (Rattus norvegicus, rat, 656, 10); mesau (Mesosorex auratus, hamster, 101, 5); galga (Gallus gallus, chicken, 297, 10); and susc (Sus scrofa, pig, 142, 10). Fungi: oryct (Oryctolagus cuniculus, rabbit, 165, 5); musmu (Mus musculus, mouse, 1,065, 50); ratio (Rattus norvegicus, rat, 656, 10); mesau (Mesosorex auratus, hamster, 101, 5); galga (Gallus gallus, chicken, 297, 10); and susc (Sus scrofa, pig, 142, 10). Invertebrates: drste (Drosophila melanogaster, sea urchin, 127, 3); dicdi (Dictyostelium discoideum, silkworm, 140, 4); and caeel (Caenorhabditis elegans, nematode, 117, 3). Plants: orycu (Oryctolagus cuniculus, rabbit, 165, 5); musmu (Mus musculus, mouse, 1,065, 50); ratio (Rattus norvegicus, rat, 656, 10); mesau (Mesosorex auratus, hamster, 101, 5); galga (Gallus gallus, chicken, 297, 10); and susc (Sus scrofa, pig, 142, 10). Each sequence collection was cleaned by removing duplicate copies of the same gene as identified in the sequence annotations. Species names are abbreviated as follows: Vertebrates: homsa (Homo sapiens, human), aggregate sequence length 3,326 kb, minimum contig size 80 kb; bost (Bos taurus, bovine, 136, 10); sucss (Sus scrofa, pig, 142, 10); rorc (Rattus norvegicus, rat, 656, 10); and musm (Mus musculus, mouse, 1,065, 50); ratio (Rattus norvegicus, rat, 656, 10); mesau (Mesosorex auratus, hamster, 101, 5); galga (Gallus gallus, chicken, 297, 10); and susc (Sus scrofa, pig, 142, 10). Fungi: oryct (Oryctolagus cuniculus, rabbit, 165, 5); musmu (Mus musculus, mouse, 1,065, 50); ratio (Rattus norvegicus, rat, 656, 10); mesau (Mesosorex auratus, hamster, 101, 5); galga (Gallus gallus, chicken, 297, 10); and susc (Sus scrofa, pig, 142, 10). Invertebrates: drste (Drosophila melanogaster, sea urchin, 127, 3); dicdi (Dictyostelium discoideum, silkworm, 140, 4); and caeel (Caenorhabditis elegans, nematode, 117, 3). Plants: orycu (Oryctolagus cuniculus, rabbit, 165, 5); musmu (Mus musculus, mouse, 1,065, 50); ratio (Rattus norvegicus, rat, 656, 10); mesau (Mesosorex auratus, hamster, 101, 5); galga (Gallus gallus, chicken, 297, 10); and susc (Sus scrofa, pig, 142, 10). Each sequence collection was cleaned by removing duplicate copies of the same gene as identified in the sequence annotations. Species names are abbreviated as follows: Vertebrates: homsa (Homo sapiens, human), aggregate sequence length 3,326 kb, minimum contig size 80 kb; bost (Bos taurus, bovine, 136, 10); sucss (Sus scrofa, pig, 142, 10); rorc (Rattus norvegicus, rat, 656, 10); and musm (Mus musculus, mouse, 1,065, 50); ratio (Rattus norvegicus, rat, 656, 10); mesau (Mesosorex auratus, hamster, 101, 5); galga (Gallus gallus, chicken, 297, 10); and susc (Sus scrofa, pig, 142, 10).
(iii) The dinucleotides CC/GG, TG/CA, and AG/CT, all a single base mutation from CG, are (except for dicot plants) only overrepresented in genomes with strong CG suppression. Interestingly, scrutiny of Table 1 reveals that these dinucleotide relative abundances separate rodents possessing TG/CA and AG/CT of significantly high representations and CC/GG in the normal range from the nonrodents (primates, artiodactyls, and lagomorphs) that possess relative high abundances of CC/GG, but TG/CA and AG/CT in the normal range.

(iv) Other dinucleotide biases in eukaryotes include over-representation of GC in Drosophila species but apparently not in other higher eukaryotes. GC is significantly abundant in most γ-proteobacteria (8, 9).

(v) No dinucleotide extremes were found in the moth B. mori or in barley (H. vulgare). Protists form a diverse group with no consistent pattern of dinucleotide relative abundances.

### Dinucleotide Relative Abundance \( \delta^\ast \)-Differences Among Eukaryotes

The sequence collections were organized into distinct \( \approx 50 \text{ kb} \) sequence samples. The average \( \delta^\ast \)-differences between samples of the same genome (within-species \( \delta^\ast \)) or from two different genomes (between-species \( \delta^\ast \)) are exhibited in Table 2. It is useful to distinguish distinct levels of \( \delta^\ast \)-differences like random \((\delta^\ast < 0.018)\), very close \((0.020 < \delta^\ast < 0.030)\), close \((0.035 < \delta^\ast < 0.050)\), moderately similar \((0.055 < \delta^\ast < 0.075)\), weakly similar \((0.080 < \delta^\ast < 0.100)\), distinctly similar \((0.120 < \delta^\ast < 0.150)\), distant \((0.160 < \delta^\ast < 0.200)\), and very distant \((\delta^\ast > 0.200)\); cf. ref. 8).

The following relations are evident in Table 2.

#### Table 2. Average \( \delta^\ast \)-differences (multiplied by 1,000) within and between eukaryotes based on \( \approx 50 \text{ kb} \) sequence samples

<table>
<thead>
<tr>
<th>Sample</th>
<th>Average ( \delta^\ast )-differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homeo</td>
<td>1.20</td>
</tr>
<tr>
<td>Drosophila</td>
<td>1.10</td>
</tr>
<tr>
<td>Rodent</td>
<td>1.00</td>
</tr>
<tr>
<td>Bird</td>
<td>0.90</td>
</tr>
<tr>
<td>Amphibian</td>
<td>0.80</td>
</tr>
<tr>
<td>Insect</td>
<td>0.70</td>
</tr>
<tr>
<td>Worm</td>
<td>0.60</td>
</tr>
<tr>
<td>Protozoa</td>
<td>0.50</td>
</tr>
<tr>
<td>Fungi</td>
<td>0.40</td>
</tr>
<tr>
<td>Animal</td>
<td>0.30</td>
</tr>
<tr>
<td>Plant</td>
<td>0.20</td>
</tr>
</tbody>
</table>

See legend to Table 1 for the list and sizes of the sequence samples.
and nonrodents (human, artiodactyls, and rabbit). The mutual signature but also accommodates amino acid constraints (10).

of genes (50 or more), we found that the codon signature, like

\[
y' \in [0.071, 0.118]
\]

Interestingly, the samples.

\[
y \in [1.20, 1.26]
\]

dinucleotide relative abundances of TG and AG \( \{2,3\} \), and \( \{3,4\} \) (4

\[
5 \in [0.41, 0.79]
\]

\( \rho_{CC} \) from 0.58 to 0.79 and marginally to significantly overrepresented in CC/GG at the level \( \rho_{CCGG} = 1.20 \pm 1.26 \). The dinucleotide relative abundances of TG/CA and AG/CT are also high normal to significantly high. In ref. 10 we introduced the codon signature, defined as the dinucleotide relative abundances at the distinct codon positions (1,2), (2,3), and (3,4) (\( \{4=1\) of the next codon). For large collections of genes (50 or more), we found that the codon signature, like the genome signature, is essentially invariant. Moreover, the codon signature in mammals largely parallels the genome signature but also accommodates amino acid constraints (10).

The available mammals split into two groups: rodents and nonrodents (human, artiodactyls, and rabbit). The mutual \( \delta^* \)-differences within each group are close but \( \delta^* \)-differences between rodents and nonrodent mammals show only moderate similarity, with hamster somewhat more distant from the nonrodents. The chicken and X. laevis sequences are weakly or moderately similar to mammals and the sea urchin is weakly similar. In fact, chicken is somewhat closer to rodents than to nonrodent mammals, whereas X. laevis is somewhat closer to nonrodent mammals than to rodents.

Insects form a diverse group with mutual \( \delta^* \)-differences in the range 0.071–0.118 (weakly similar). Interestingly, the insects tend to be weakly similar to monocots but generally distantly similar to dicots.

(v) Fungi constitute a coherent group with most \( \delta^* \)-differences in the range from 0.035 to 0.075, close or moderately similar. Exceptions are C. albicans and the smut U. maydis, both distant from other fungi. The \( \delta^* \)-differences between the 16 chromosomes of the yeast S. cerevisiae are remarkably close. The within-chromosome \( \delta^* \)-differences range is 0.018 to 0.024 except for chromosome 1 (average \( \delta^* \), 0.035), and the between-chromosome range is 0.018 to 0.030, indicating that all chromosomes are very close to each other (see Fig. 1). \( \delta^* \)-differences within and between chromosomes II and III of C. elegans are remarkably similar with mean \( \delta^* \)-differences about 0.038, marginally higher than the within whole yeast genome.

(vi) In the group of plants, the three dicot species (tobacco, potato, and tomato, all in the Solanaceae family) are very close (0.027–0.032) and their \( \delta^* \)-differences to the dicot A. thaliana are at the level of moderate similarity (0.058–0.063). Our samples from monocots are mutually very close or close and only moderately similar to dicots. Plants (equally with respect to both dicots and monocots) are close to S. cerevisiae (\( \delta^* \)-differences 0.036–0.054) and also to the yeast-like K. lactis, but only moderately or weakly similar to other fungi. Thus, in terms of DNA normalized doublet comparisons of whole genomes, fungi appear to be closer to plants than to animals, a result that is at variance with some protein sequence comparisons (11). Dicot plants are weakly similar to mammals (\( \delta^* \)-differences 0.070–0.132). On the other hand, insects are distant or very distant to mammals.

(vii) Protists are a diverse group. All \( \delta^* \)-differences from P. fulciparum to other eukaryotes exceed 0.090 and mostly exceed 0.120. T. brucei is moderately similar to some insects [the closest \( \delta^*(trybr,bommo) = 0.068 \), some fungi [\( \delta^*(trybr,schpo) = 0.059 \)], and monocot plants (\( \delta^* \)-differences 0.058–0.062). D. discoideum is weakly similar to yeast and dicot plants.

Comparisons of mt Genomes Between Species

The \( \delta^* \)-differences between various mt-genomes and separately between the nuclear genomes of the corresponding host species are given in Table 3. The principal observation is that levels of similarity among mt genomes assessed by \( \delta^* \)-differences largely parallel levels of evolutionary relatedness assessed by the genome signature \( \delta^* \)-differences among the corresponding host genomes (Fig. 2). That is, species exhibiting small (large) \( \delta^* \)-differences between their nuclear genomes generally have relatively small (large) \( \delta^* \)-differences between corresponding mt genomes and vice versa. However, comparing the mt genomic signature with its host genome signature shows \( \delta^* \)-differences generally distant to very distant (\( \delta^* \) values mostly 0.130–0.230, see Table 3) with no discernable pattern among the various eukaryotic kingdoms.

Organellar genomes (mitochondria and chloroplasts) are widely accepted as bacterial endosymbionts in that these genomes constitute the remnants of once free-living cellular organisms (12). There is great diversity among mitochondria, including substantial size variation and contrasting patterns of mt genome organization and gene expression relative to animal, plant, fungal, and protist lineages (12). A central unresolved problem concerns whether mt evolution (primary and/or secondary endosymbiont events) is monophyletic or polyphyletic.

The reported D. melanogaster mt genome (19,517 bp) contains about 4 kb (15,566-end) consisting of 12 copies, each about 350 bp long, of the NADH–ubiquinone oxidoreductase chain 6 of unusual composition. This 4 kb section was removed and the remaining sequence used as the adjusted D. melanogaster mt genome. The S. cerevisiae mtDNA (∼78 kb) composition is an extreme anomaly. This is attested to by the

![Fig. 1. Distribution of \( \delta^* \)-differences within and between selected species based on all pairwise comparisons of ∼50-kb disjoint sequence samples.](image)
exorbitant $\delta^*$-differences from all other mt or nuclear genomes (Table 3), in part due to the more than 100 G+C-rich clusters, each about 50 bp long, separated by A+T-rich spacer regions and to the numerous transposable elements. The $\delta^*$-differences (Table 3) reflected in the tree (Fig. 2) place Trypanosoma far out. It is possible that the mt–Trypanosoma endosymbiont reflects a distinct event from that of the other organisms or, like the S. cerevisiae endosymbiont, possesses highly anomalous nucleotide organization due to invasion by transposons or other drastic mutational events.

Animal (vertebrate and invertebrate) mt sequences show significant underrepresentations of CG dinucleotides, $p_{CG} \approx 0.40$ to 0.60 (13), almost to the same extent as occurs in vertebrate genome sequences. The adjusted $D. melanogaster$ and (unadjusted) $D. yakuba$ mt genomes entail $p_{CG} = 0.73$ and 0.68, respectively (13). The fungal S. pombe has $p_{CG} = 0.54$ typical of animal mitochondria. However, the Podospora anserina fungal mt CG representation is in the normal range. The mt genome of A. thaliana has $p_{CG} = 0.73$, significantly low. The single persistent significantly high $p^*$ value occurs for $p_{CC/GG}$ typical in animal and fungal mt sequences. Intriguingly, the chloroplast genomes are all significantly high in $p_{CC/GG}$, which is the only consistent extreme among currently available chloroplast sequences (5).

Molecular Evolutionary Implications

Based on comparisons of both the dinucleotide relative abundance extremes (genome signature) and $\delta^*$-differences we venture some interpretations of molecular evolutionary relationships among eukaryotic nuclear and mt genomes. In this context, we hypothesize that specificity in replication and repair machinery and context-dependent mutation biases largely maintain the homogeneity of the whole genome of an organism as reflected in the constancy of dinucleotide relative abundances and that differences in this machinery produce the observed differences in the dinucleotide relative abundances among species (1, 9, 14).

(i) Rodents are somewhat exceptional among mammals in $\delta^*$-differences. In particular, the nonrodent mammals (human, artiodactyls, rabbit) and rodents (mouse, rat, hamster) constitute two coherent groups but show only moderate intergroup similarity. What can account for this separation? Rodents tend to have a higher mutation rate and shorter generation time than many nonrodent mammals (15). Moreover, it is established that rodents are inefficient in global repair of cyclobutane thymidine dimers compared with humans and probably also in repair of other forms of oxidative DNA damage (16). These differences relate principally to replication and repair mechanisms and context-dependent mutation tendencies, consistent with our hypothesis that such molecular differences could produce the observed differences in the genome signature (1, 9). One might inquire about the reasons for differences in repair proficiency of thymidine dimers between human and rodents. We speculate that the rodents analyzed live in more secluded (often underground or noctur-

---

**Table 3. $\delta^*$-differences between mitochondrial and the host nuclear genomes (multiplied by 1,000)**

<table>
<thead>
<tr>
<th>homsa</th>
<th>bosta</th>
<th>musmu</th>
<th>ratno</th>
<th>galga</th>
<th>xenla</th>
<th>strpu</th>
<th>drome</th>
<th>caeel</th>
<th>sacce</th>
<th>schpo</th>
<th>arath</th>
<th>trybr</th>
</tr>
</thead>
<tbody>
<tr>
<td>145</td>
<td>25</td>
<td>26</td>
<td>33</td>
<td>36</td>
<td>50</td>
<td>101</td>
<td>139</td>
<td>85</td>
<td>468</td>
<td>76</td>
<td>158</td>
<td>184</td>
</tr>
<tr>
<td>37</td>
<td>135</td>
<td>23</td>
<td>40</td>
<td>48</td>
<td>36</td>
<td>91</td>
<td>130</td>
<td>91</td>
<td>457</td>
<td>68</td>
<td>145</td>
<td>188</td>
</tr>
<tr>
<td>56</td>
<td>55</td>
<td>153</td>
<td>17</td>
<td>34</td>
<td>37</td>
<td>96</td>
<td>120</td>
<td>87</td>
<td>457</td>
<td>72</td>
<td>154</td>
<td>177</td>
</tr>
<tr>
<td>56</td>
<td>52</td>
<td>40</td>
<td>146</td>
<td>26</td>
<td>44</td>
<td>97</td>
<td>112</td>
<td>88</td>
<td>461</td>
<td>72</td>
<td>154</td>
<td>174</td>
</tr>
<tr>
<td>69</td>
<td>72</td>
<td>59</td>
<td>59</td>
<td>165</td>
<td>52</td>
<td>93</td>
<td>130</td>
<td>101</td>
<td>487</td>
<td>68</td>
<td>150</td>
<td>183</td>
</tr>
<tr>
<td>61</td>
<td>55</td>
<td>77</td>
<td>75</td>
<td>88</td>
<td>85</td>
<td>81</td>
<td>106</td>
<td>81</td>
<td>465</td>
<td>66</td>
<td>124</td>
<td>172</td>
</tr>
<tr>
<td>110</td>
<td>93</td>
<td>108</td>
<td>98</td>
<td>116</td>
<td>89</td>
<td>172</td>
<td>116</td>
<td>98</td>
<td>486</td>
<td>52</td>
<td>94</td>
<td>177</td>
</tr>
<tr>
<td>173</td>
<td>169</td>
<td>196</td>
<td>197</td>
<td>182</td>
<td>144</td>
<td>147</td>
<td>175</td>
<td>96</td>
<td>385</td>
<td>116</td>
<td>163</td>
<td>167</td>
</tr>
<tr>
<td>169</td>
<td>153</td>
<td>183</td>
<td>177</td>
<td>172</td>
<td>156</td>
<td>117</td>
<td>99</td>
<td>228</td>
<td>439</td>
<td>92</td>
<td>168</td>
<td>131</td>
</tr>
<tr>
<td>126</td>
<td>109</td>
<td>138</td>
<td>136</td>
<td>133</td>
<td>98</td>
<td>78</td>
<td>91</td>
<td>75</td>
<td>527</td>
<td>496</td>
<td>542</td>
<td>388</td>
</tr>
<tr>
<td>142</td>
<td>131</td>
<td>157</td>
<td>156</td>
<td>142</td>
<td>113</td>
<td>105</td>
<td>77</td>
<td>72</td>
<td>36</td>
<td>130</td>
<td>84</td>
<td>188</td>
</tr>
<tr>
<td>124</td>
<td>103</td>
<td>131</td>
<td>125</td>
<td>132</td>
<td>100</td>
<td>66</td>
<td>122</td>
<td>78</td>
<td>42</td>
<td>59</td>
<td>83</td>
<td>240</td>
</tr>
<tr>
<td>154</td>
<td>147</td>
<td>165</td>
<td>165</td>
<td>152</td>
<td>124</td>
<td>107</td>
<td>90</td>
<td>92</td>
<td>67</td>
<td>59</td>
<td>84</td>
<td>219</td>
</tr>
</tbody>
</table>

*Included are all eukaryotes for which both the complete mitochondrial genome and a sufficient sample of nuclear genomic DNA is available.

$\delta^*$-differences between mitochondrial and the host nuclear genomes (multiplied by 1,000) in the lower right triangle, $\delta^*$-difference between the mitochondrial genome and the corresponding host (boldface) are shown in the diagonal, and average $\delta^*$-difference between corresponding nuclear genomes (italics) in the lower left triangle.

---

**Fig. 2.** Evolutionary tree derived from $\delta^*$-differences between mt (Upper) and nuclear (Lower) genomes. The trees were generated by the UPGMA (average linkage) method. The scale ($\delta^*$-differences multiplied by 1,000) is shown at the bottom.
tional habitats with lesser exposure to sunlight and other sources of radiation damage. In these environments, we could expect that natural selection has attenuated relevant molecular mechanisms required for efficient repair of cyclobutane dimer adducts. However, embryonic rodent and human cells do not show differences in repair processes and transcription-coupled repair processes are largely similar in humans and rodents (16). The relative abundance value of TT/AA in mouse and rat sequences is about 1.06 (1.03 in hamster) compared with 1.12–1.14 in human, cow, pig, and rabbit. On this basis, there are fewer occurrences of TT (thymidine dimer) in rodents compared with nonrodent mammals.

The foregoing discussion suggests a testable hypothesis. There are rodents active in relatively exposed environments, including the grey squirrel and especially capybara. In these cases, one could expect an efficient cyclobutane dimer repair system. On the other hand, purely nocturnal mammals (e.g., the owl monkey) are conceivably inefficient in thymidine dimer adduct repair.

(ii) δ+*-differences in the dicots (tobacco, potato, tomato) are mutually close and similarly for the monocots (maize, barley, rice), whereas the monocot and these dicot sequences are only moderately similar (between-group δ+*-differences larger by a factor of about two than within-group δ+*-differences). Thus, the fundamental monocot/dicot morphological distinction parallels a significant genome signature difference. Intriguingly, the fungi are equally distant to monocots and dicots, whereas the insects are closer to monocots over dicots by a factor of about 1.4. The separation of dicots and monocots (the grasses considered to have arisen from a subgroup of dicots) may have occurred about 200 million years ago (17).

(iii) The genome signature comparisons are in agreement with the classic division of most metazoan phyla into the deuterostomes (e.g., vertebrates, echinoderms) and protostomes (insects, worms), since the sea urchin is weakly similar to vertebrates but the protostomes are distant or very distant from the vertebrates.

(iv) Compositional biases across bacterial genomes were discussed in Karlin et al. (9). The dinucleotide relative abundance values (genomic signature) of bacterial genomes place the *Sulfolobus*-like sequences (eocyte phylogeny) closer to vertebrates than are all other bacterial genomes, and cyanobacteria closer to fungi and plants (4, 8). Enigmatically *Haemophilus influenzae* is moderately similar to *D. melanogaster* sequences.

(v) A challenging question concerns reasons and mechanisms to account for the qualitative concordance between the evolutionary development of host nuclear genomes and the development of mt organelle genomes despite the pronounced difference between the mt and host nuclear genome signatures. The mt and nuclear genomes for animal and fungal organisms use independent DNA polymerase machinery (e.g., γ vs. α, e, δ subunits, respectively). Also, the methods of replication and the nature of the replication origins are fundamentally different. Specifically, the animal and fungal mt transcription-primed replication machinery is distinctive in that most of the heavy strand is synthesized first and the light strand subsequently, whereas the nuclear genomes are replicated analogously to eubacteria synchronized over multiple replication origins.

What about influences of repair processes? There appears to be no DNA excision repair mechanism to deal with cyclobutane dimers in the mitochondrion and apparently bulky lesions are not repaired (18). mtDNA in animals and fungi shows elevated levels of single- and double-strand breaks, mismatches, and generally corrupted base pairings (19). This may be due to a paucity of abasic site correction facilities and mismatch repair capacity in mt genomes (19). Moreover, repair may be less urgent for mt activity because each cell has many mitochondria (hundreds or thousands) and a modicum of impaired organelles may not significantly curtail energy production.

Notably, virtually all mitochondria maintain normal representations of TA dinucleotides, whereas nuclear DNA overwhelmingly tends to have TA in low relative abundance, suggesting that mtDNA may be less thermodynamically stable than nuclear DNA because the dinucleotide TA has the lowest stacking energies compared with all other base steps (6). We thank Drs. B. E. Blaisdell, A. M. Campbell, and C. Burge for helpful comments on the manuscript. This work was supported in part by National Institutes of Health Grants 2R01GM10452–32 and 5R01HG00335–09, and National Science Foundation Grant 9403553–002.