Subfamily Relationships and Clustering of Rabbit C Repeats

Dan E. Krane,* Andrew G. Clark,† Jan-Fang Cheng,*2 and Ross C. Hardison*

*Department of Molecular and Cell Biology and †Department of Biology, The Pennsylvania State University

C repeats constitute the predominant family of short interspersed repeats (SINES) in the rabbit genome. Determination of the nucleotide sequence 5’ to rabbit ζ-globin genes reveals clusters of C repeats, and analysis of these and other sequenced regions of rabbit chromosomes shows that the C repeats have a strong tendency to insert within or in close proximity to other C repeats. An alignment of 44 members of the C repeat family shows that they are composites of different sequences, including a tRNA-like sequence, a conserved central core, a stretch of repeating CT dinucleotides, and an A-rich tract. Cladograms generated by both parsimony and cluster analysis subdivide the C repeats into at least three distinct subfamilies. Nucleotides at sites diagnostic for subfamilies appear to have changed in a punctuated and progressive manner during evolution, indicating that a limited number of progenitors have given rise to new repeats in waves of dispersion. C repeats that insert into preexisting C repeats belong to subfamilies that are proposed to have been propagated more recently; hence, these data support the model of dispersion in successive waves. The divergence among the oldest group of C repeats is greater than that observed for the analogous Alu repeats in humans, indicating that rabbit C repeats have been propagating longer than human Alu repeats. The improved consensus sequence for these repeats is similar to that of the predominant artiodactyl SINE in both the tRNA-like region and a central region. Because members of different subfamilies cross-hybridize very poorly, hybridization data with representatives of each subfamily provide a new minimal estimate, 234,000, for the copy number of C repeats in the rabbit haploid genome, although it is likely that the actual value is closer to 1 million.

Introduction

Multiple copies of different repetitive sequences interspersed throughout a eukaryotic genome contribute significantly to the size and possibly to the function of that genome (Britten and Kohne 1968; Davidson and Britten 1973). In mammals, these repetitive elements have been classified by length as either SINES (short), typically <500 bp in length, or LINES (long) with full-length copies ranging in size from 6 to 8 kb (Singer 1982). Both size classes are proposed to be propagated by retrotransposition (Weiner et al. 1986), a process by which RNAs transcribed from these repeats

1. Key words: C repeats, short interspersed repeats, transposable elements, genome evolution, tRNA, copy number, ζ-globin genes, DNA insertions, clustering.
2. Current address: Human Genome Center, 459 Donner Laboratory, Lawrence Berkeley Laboratory, Berkeley, California 94720.

Address for correspondence and reprints: Dr. Ross C. Hardison, Department of Molecular and Cell Biology, The Pennsylvania State University, University Park, Pennsylvania 16802.
or their source genes are reverse-transcribed into DNA and reinserted into the genome at staggered breaks, repair of which produces flanking direct repeats. Although widely dispersed in the human \textit{(Homo sapiens)} genome, the predominant LINE (L1 repeat) tends to be found in A+T-rich G bands in chromosomes, while the predominant SINE \textit{(Alu} repeat) is found in G+C-rich R bands (Korenberg and Rykowski 1988), indicating a strong regional preference for stable insertion.

Each mammalian order has one or more characteristic families of SINES that are present in high copy number (e.g., primate \textit{Alu}, lagomorph \textit{C}, rodent \textit{B1} and \textit{B2}, and artiodactyl “consensus repeats”). However, sequence comparisons show that SINES in each order have different origins (Schmid and Shen 1985). For example, human \textit{Alu} repeats probably arose from a 7SL RNA gene (Ullu and Tschudi 1984), while many SINES in other mammalian orders, including rabbit \textit{(Oryctolagus cuniculus)} \textit{C} repeats, appear to be derived in part from tRNA genes (Daniels and Deininger 1985; Kazuichi and Okada 1985; Lawrence et al. 1985). Despite the sequence differences, several structural features are shared by most mammalian SINES, including internal promoters for RNA polymerase III, A-rich tracts at their 3' ends, a size of \(\sim 300\) nucleotides, and an absence of conserved open reading frames (Weiner et al. 1986). Several polymorphisms are explained by the recent insertion of SINES (Economou-Pachnis and Tsichlis 1985; Friezner-Degen et al. 1986; King et al. 1986; Stoppa-Lyonnet et al. 1990), indicating that some of these repeats are still actively propagating.

\textit{Alu} repeats, which are the predominant SINE in humans, can be divided into at least three (Willard et al. 1987) and probably four (Britten et al. 1988) subfamilies that are apparently of different ages. Divergence among members of the subfamilies increases with age (Willard et al. 1987). Nucleotide positions diagnostic for each subfamily have been identified (Britten et al. 1988), and the sequences of older repeats are progressively more like the probable 7SL RNA ancestor (Britten et al. 1988; Jurka and Smith 1988). The ability to separate these repeats into well-defined subfamilies argues for a model in which only a very small number of sequences, progenitor repeats, are capable of generating new repeats that can insert into the genome (Britten et al. 1988). A similar pattern of subfamilies is also seen in the rodent family of \textit{B2}, but not \textit{B1}, repeats (Bains and Temple-Smith 1989).

\textit{Ah} repeats, which are the predominant SINE in humans, can be divided into at least three (Willard et al. 1987) and probably four (Britten et al. 1988) subfamilies that are apparently of different ages. Divergence among members of the subfamilies increases with age (Willard et al. 1987). Nucleotide positions diagnostic for each subfamily have been identified (Britten et al. 1988), and the sequences of older repeats are progressively more like the probable 7SL RNA ancestor (Britten et al. 1988; Jurka and Smith 1988). The ability to separate these repeats into well-defined subfamilies argues for a model in which only a very small number of sequences, progenitor repeats, are capable of generating new repeats that can insert into the genome (Britten et al. 1988). A similar pattern of subfamilies is also seen in the rodent family of \textit{B2}, but not \textit{B1}, repeats (Bains and Temple-Smith 1989).

\textit{C} repeats are typical SINES. With approximately 1 million copies interspersed throughout the genome (present paper) and an average length of 324 bp (Cheng et al. 1984; Hardison and Printz 1985), they account for \(\sim 13\)% of the genomic DNA in rabbits. Several individual \textit{C} repeats contribute sequences useful to rabbits by providing functional polyadenylation signals to the genes into which they have inserted (Krane and Hardison 1990). In the present study new sequence data from the rabbit \(\alpha\)-like globin gene cluster reveals two large clusters of \textit{C} repeats. Alignment of these and other sequenced \textit{C} repeats shows that they fall into at least three distinct subfamilies whose progenitors have diagnostic sites that have changed in a progressive and apparently punctuated manner, much like human \textit{Alu} and mouse \textit{B2} repeats. In clusters of \textit{C} repeats formed by recursive integrations, members of the more recent families insert into members of the older families, thus independently confirming the model of waves of dispersion.

**Material and Methods**

**Nucleotide Sequence Determination**

Clones of rabbit DNA containing \(\xi\)-, \(\alpha\)-, and \(\theta\)-globin genes (the \(\alpha\)-like globin gene cluster) were previously isolated from a library of rabbit genomic DNA (Cheng
The regions 5' to the ζ1 and ζ2 genes containing several C repeats were sequenced mainly by the dideoxynucleotide chain-termination method of Sanger et al. (1977). For the sequence 5' to gene ζ1, restriction fragments were subcloned into M13mp18 and M13mp19 to generate templates for sequencing. Subclones containing progressive deletions of the ζ2 gene were constructed using exonuclease III and mung-bean nuclease (Henikoff 1984), to provide templates for sequencing the cluster of C repeats 5' to the ζ2 gene. The sequence of that region was confirmed in part by the chemical degradation method of Maxam and Gilbert (1977). Maps of these regions have been presented by Cheng et al. (1988).

Alignment of Multiple Sequences

A multiple alignment of 28 C repeat sequences (C1–C25 and C28–C30 in table 1) was made by first generating two-way alignments between the repeats by the Genetic Computer Group's Gap program (Devereux et al. 1984), which uses the alignment algorithm of Needleman and Wunsch (1970). These two-way alignments were compared and aligned by inspection, and gaps were inserted to increase the average pairwise similarity. Further improvements were then made by a reiterative process of cladogram generation and realignment based on the groupings, to improve the alignment of outliers. An initial consensus sequence based on the alignment of these 28 repeats was then generated. Unambiguous assignments were made either when >60% of the nucleotides at a given site were the same or when the frequency of any single nucleotide was at least twofold greater than the next most frequently occurring nucleotide. At the remaining positions, twofold ambiguous assignments (e.g., purine) were made when the total percentage of two nucleotides at a site exceeded 67%, threefold ambiguous assignments (e.g., not G) were made when the total percentage of three nucleotides at a site was >89%, and all other sites were given the fourfold ambiguous assignment (N or “any nucleotide”). No nucleotides were included in the consensus when fewer than one-third of the repeats had a nucleotide at that site in the multiple alignment.

This initial consensus was used to search the GenBank data base (release 60.0, June 1989) with the WordSearch (Devereux et al. 1984) and FASTA (Pearson and Lipman 1988) computer programs. Twelve previously unrecognized repeats found in this way (C26, C27, C31–C34, and C39–C44 in Table 1) were added to the growing multiple alignment as they were discovered. The current consensus is based on the 44 repeats listed in table 1. Similarities between the consensus sequence and individual repeats were determined with the Gap program, which treats matches between unique and ambiguous nucleotides equally. Match scores involving some ambiguous positions are referred to as percentage similarities, while matches between sequences with only unambiguous nucleotides are called percentage identities.

Phylogenetic Analysis

The sequences were analyzed phylogenetically by two approaches based on different principles. In the first, a dendrogram was constructed to provide a graphical representation of the matrix of pairwise identities between C repeats. The identity matrix was constructed using the Gap program, with a gap penalty of 4.5 and a gap-length penalty of 0.3/nucleotide. Grossly truncated repeats (C2, C6, C25, C27, and C33) were excluded from this cluster analysis. An average-distance dendrogram was constructed using the cluster algorithm P1M of the package BMDP (Hartigan 1983).

In the second approach, phylograms were generated from the sequences on the basis of the principle of maximum parsimony and were verified by bootstrapping
Table 1
Sources of C Repeat Sequences Used in Present Study

<table>
<thead>
<tr>
<th>GenBank Accession</th>
<th>Repeat Number</th>
<th>Start*</th>
<th>End*</th>
<th>Orientation</th>
<th>Location</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X07786</td>
<td>C1</td>
<td>2777</td>
<td>3245</td>
<td>-</td>
<td>5' to α globin</td>
<td>Margot et al. 1989</td>
</tr>
<tr>
<td>X07786</td>
<td>C2</td>
<td>3552</td>
<td>3827</td>
<td>+</td>
<td>5' to α globin</td>
<td>Cheng et al. 1984; Margot et al. 1989</td>
</tr>
<tr>
<td>X07786</td>
<td>C3</td>
<td>8894</td>
<td>9206</td>
<td>-</td>
<td>5' to γ globin</td>
<td>Margot et al. 1989</td>
</tr>
<tr>
<td>X07786</td>
<td>C4</td>
<td>10035</td>
<td>10422</td>
<td>-</td>
<td>5' to γ globin</td>
<td>Margot et al. 1989</td>
</tr>
<tr>
<td>X07786</td>
<td>C5</td>
<td>10444</td>
<td>10809</td>
<td>-</td>
<td>5' to γ globin</td>
<td>Margot et al. 1989</td>
</tr>
<tr>
<td>X07786</td>
<td>C6</td>
<td>10866</td>
<td>11186</td>
<td>-</td>
<td>5' to γ globin</td>
<td>Margot et al. 1989</td>
</tr>
<tr>
<td>X07786</td>
<td>C7</td>
<td>11814</td>
<td>12184</td>
<td>-</td>
<td>5' to γ globin</td>
<td>Margot et al. 1989</td>
</tr>
<tr>
<td>X07786</td>
<td>C8</td>
<td>18604</td>
<td>18936</td>
<td>+</td>
<td>5' to δ globin</td>
<td>Cheng et al. 1984; Margot et al. 1989</td>
</tr>
<tr>
<td>X07786</td>
<td>C9</td>
<td>18437</td>
<td>19085</td>
<td>+</td>
<td>5' to δ globin</td>
<td>Margot et al. 1989</td>
</tr>
<tr>
<td>X07786</td>
<td>C10</td>
<td>19133</td>
<td>19470</td>
<td>+</td>
<td>5' to δ globin</td>
<td>Margot et al. 1989</td>
</tr>
<tr>
<td>X07786</td>
<td>C11</td>
<td>20025</td>
<td>20744</td>
<td>-</td>
<td>5' to δ globin</td>
<td>Margot et al. 1989</td>
</tr>
<tr>
<td>X07786</td>
<td>C12</td>
<td>20381</td>
<td>20743</td>
<td>-</td>
<td>5' to δ globin</td>
<td>Margot et al. 1989</td>
</tr>
<tr>
<td>X07786</td>
<td>C13</td>
<td>24175</td>
<td>24573</td>
<td>-</td>
<td>5' to δ globin</td>
<td>Margot et al. 1989</td>
</tr>
<tr>
<td>X07786</td>
<td>C14</td>
<td>26206</td>
<td>26586</td>
<td>+</td>
<td>5' to β globin</td>
<td>Cheng et al. 1984; Margot et al. 1989</td>
</tr>
<tr>
<td>X07786</td>
<td>C15</td>
<td>32897</td>
<td>33182</td>
<td>-</td>
<td>3' to β globin</td>
<td>Margot et al. 1989</td>
</tr>
<tr>
<td>M35026</td>
<td>C16</td>
<td>348</td>
<td>700</td>
<td>+</td>
<td>5' to η globin</td>
<td>Present paper</td>
</tr>
<tr>
<td>M35026</td>
<td>C17</td>
<td>332</td>
<td>975</td>
<td>+</td>
<td>5' to η globin</td>
<td>Present paper</td>
</tr>
<tr>
<td>M35026</td>
<td>C18</td>
<td>994</td>
<td>1996</td>
<td>+</td>
<td>5' to η globin</td>
<td>Present paper</td>
</tr>
<tr>
<td>M35026</td>
<td>C19</td>
<td>1311</td>
<td>1655</td>
<td>-</td>
<td>5' to η globin</td>
<td>Present paper</td>
</tr>
<tr>
<td>M35026</td>
<td>C20</td>
<td>1304</td>
<td>1994</td>
<td>+</td>
<td>5' to η globin</td>
<td>Present paper</td>
</tr>
<tr>
<td>M35026</td>
<td>C21</td>
<td>3888</td>
<td>4159</td>
<td>-</td>
<td>5' to α globin</td>
<td>Present paper</td>
</tr>
<tr>
<td>M35026</td>
<td>C22</td>
<td>5341</td>
<td>5729</td>
<td>+</td>
<td>5' to α globin</td>
<td>Present paper</td>
</tr>
<tr>
<td>J03884</td>
<td>C23</td>
<td>1Δ</td>
<td>263Δ</td>
<td>+</td>
<td>5' to η globin</td>
<td>Present paper</td>
</tr>
<tr>
<td>J03884</td>
<td>C24</td>
<td>257</td>
<td>588</td>
<td>+</td>
<td>5' to η globin</td>
<td>Present paper</td>
</tr>
<tr>
<td>J03884</td>
<td>C25</td>
<td>775</td>
<td>1067</td>
<td>-</td>
<td>5' to η globin</td>
<td>Present paper</td>
</tr>
<tr>
<td>J00687</td>
<td>C26</td>
<td>240</td>
<td>546</td>
<td>+</td>
<td>5' to uteroglobin</td>
<td>Suske et al. 1983</td>
</tr>
<tr>
<td>J00687</td>
<td>C27</td>
<td>1Δ</td>
<td>200Δ</td>
<td>+</td>
<td>5' to uteroglobin</td>
<td>Suske et al. 1983</td>
</tr>
<tr>
<td>J00687</td>
<td>C28</td>
<td>4347</td>
<td>4698</td>
<td>+</td>
<td>Intron 1 of uteroglobin</td>
<td>Suske et al. 1983; Cheng et al. 1984</td>
</tr>
<tr>
<td>M23842</td>
<td>C29</td>
<td>211</td>
<td>618</td>
<td>+</td>
<td>Unknown</td>
<td>Cheng et al. 1984</td>
</tr>
<tr>
<td>M23843</td>
<td>C30</td>
<td>161</td>
<td>570</td>
<td>+</td>
<td>Unknown</td>
<td>Cheng et al. 1984</td>
</tr>
<tr>
<td>M12280</td>
<td>C31</td>
<td>73Δ</td>
<td>303Δ</td>
<td>+</td>
<td>P-450 “exon I”</td>
<td>Zaphiropoulos et al. 1986</td>
</tr>
<tr>
<td>M14851</td>
<td>C32</td>
<td>1070Δ</td>
<td>1992Δ</td>
<td>3' to T cell receptor</td>
<td>Kornatsu et al. 1987</td>
<td></td>
</tr>
<tr>
<td>M11728</td>
<td>C33</td>
<td>1430Δ</td>
<td>1746Δ</td>
<td>3' UT of P-450 iso 4</td>
<td>Okino et al. 1985; Krane and Hardison 1990</td>
<td></td>
</tr>
<tr>
<td>J03542</td>
<td>C34</td>
<td>2555Δ</td>
<td>2904Δ</td>
<td>3' UT of apoprotein</td>
<td>Boggaram et al. 1988; Krane and Hardison 1990</td>
<td></td>
</tr>
<tr>
<td>M17431</td>
<td>C35</td>
<td>1551Δ</td>
<td>193Δ</td>
<td>3' UT of MHC gene</td>
<td>Reiher et al. 1987; Krane and Hardison 1990</td>
<td></td>
</tr>
<tr>
<td>M35026</td>
<td>C36</td>
<td>4960</td>
<td>5235</td>
<td>+</td>
<td>5' to α globin, like C17</td>
<td>Present paper</td>
</tr>
<tr>
<td>M35031</td>
<td>C37</td>
<td>359</td>
<td>702</td>
<td>+</td>
<td>5' to LLOc6</td>
<td>D. Pasqualone, D. Price, and R. Hardison, unpublished results</td>
</tr>
<tr>
<td>M35031</td>
<td>C38</td>
<td>827</td>
<td>1194</td>
<td>+</td>
<td>5' to LLOc6</td>
<td>D. Pasqualone, D. Price, and R. Hardison, unpublished results</td>
</tr>
<tr>
<td>M20015</td>
<td>C39</td>
<td>1Δ</td>
<td>321Δ</td>
<td>+</td>
<td>5' Metallothionein receptor</td>
<td>Tam et al. 1988</td>
</tr>
<tr>
<td>X06623</td>
<td>C40</td>
<td>152</td>
<td>502</td>
<td>+</td>
<td>5' Progesterone receptor</td>
<td>Milgrom et al. 1988</td>
</tr>
<tr>
<td>J03744</td>
<td>C41</td>
<td>455</td>
<td>1331</td>
<td>+</td>
<td>3' UT of myelin</td>
<td>Naravanan et al. 1988</td>
</tr>
<tr>
<td>M19707</td>
<td>C42</td>
<td>400</td>
<td>744</td>
<td>-</td>
<td>Intron 3 of MHC</td>
<td>Sittisombut et al. 1988</td>
</tr>
<tr>
<td>M19707</td>
<td>C43</td>
<td>1474</td>
<td>1790</td>
<td>-</td>
<td>Intron 5 of MHC</td>
<td>Sittisombut et al. 1988</td>
</tr>
<tr>
<td>M19707</td>
<td>C44</td>
<td>2067</td>
<td>2446</td>
<td>+</td>
<td>3' UT of MHC</td>
<td>Sittisombut et al. 1988</td>
</tr>
<tr>
<td>Y00347</td>
<td>Hare15</td>
<td>3408Δ</td>
<td>3646Δ</td>
<td>3' to hare β globin</td>
<td>Pauplin and Rech 1987</td>
<td></td>
</tr>
</tbody>
</table>

* Positions of repeats within GenBank files. A delta symbol (Δ) indicates that a repeat’s flanking direct repeat cannot be determined because of truncation of the sequence. Several repeats appear to be longer than the repeats’ average length of 324 nucleotides, because of the insertion of additional repeats.

+ A plus sign (+) indicates that the repeat is read 5' to 3', left to right on the top strand of the sequence, while a minus sign (−) indicates the opposite orientation.

+ 3'UT = 3' untranslated region; P-450 iso 4 = rabbit isozyrne 4 of cytochrome P-450; MHC = major histocompatibility complex; LLOc6 = a member of the LINE1 family of repeats in rabbits.
Subfamilies and Clustering of Rabbit C Repeats

(Swofford 1990). The PAUP program constructs trees based on the minimum number of mutational changes necessary to account for the sequences as they appear in the multiple alignment. A consensus based on bovine and goat sequences of a major family of SINES (Spence et al. 1985) was used as an outgroup in the PAUP analysis.

Hybridization Assays to Determine Copy Number

Radiolabeled probes specific to each subfamily were hybridized to slot blots containing varying amounts of purified rabbit genomic DNA or restriction fragments containing C repeats. Membranes were hybridized overnight at 45°C in a solution of 0.5 M Na phosphate, 7% Na dodecyl sulfate (SDS), 0.25 mg salmon sperm DNA/ml, adjusted to pH 7.2 (Church and Gilbert 1984). Membranes were then washed for 1 h in 1 mM ethylenediaminetetraacetic acid, 0.04 M Na phosphate, 2.5% SDS, pH 7.2, following two initial washes in 1 mM ethylenediaminetetraacetic acid, 0.04 M Na phosphate, 5% SDS, pH 7.2, at 45°C for 15 min each. The amount of hybridization was quantitated by analyzing autoradiograms with an LKB Ultroscan XL laser densitometer and was plotted against the amount of DNA. The ratio between the slopes of the lines for genomic DNA and cloned C repeats gives the fraction of the rabbit genome occupied by C repeats that hybridize to that probe, which was converted to copy number by multiplying by the size of the rabbit genome (3 x 10^9 bp) and dividing by the average length of a C repeat (324 bp).

Results

Groups of C Repeats in Rabbit α-Like Globin Gene Cluster

The α-like globin gene cluster, which was generated by block duplications of a set of genes (Cheng et al. 1987), has been the site of numerous insertions of SINES. As shown in figure 1, at least 14 C repeats are found in roughly 30 kb of this gene cluster. Sequences containing eight of these C repeats are presented in figures 2 and 3; five are located 5' to ξ1 and three are 5' to ξ2. The 11 sequenced C repeats account for 3.4 kb of 14.4 kb, or 24% of the sequenced portion, but they are not evenly distributed through the gene cluster. Instead, they tend to fall in clusters, leaving some relatively long regions (e.g., α-θ1) free of C repeats. In addition, a C repeat is found in the 5' portion of each complete J sequence in this region (fig. 1). J sequences mark junctions between homology blocks containing ξ genes (Z blocks) and θ genes (T blocks), and they have been proposed to play a role in recombination within this gene cluster (Cheng et al. 1987). Because only one C repeat in the 5' end of a J sequence has been sequenced, it is not yet clear whether the C repeat inserted prior to the duplication of the J sequences or whether these are independent insertions.

Clustering of C Repeats

Although C repeats are found throughout the rabbit genome, statistical analyses confirm that their insertion is not random and that they tend to cluster. Most of the sequenced repeats have inserted close to each other, and 11 examples of multiple insertions of repeats at single sites can be seen in the rabbit α- and β-like globin gene clusters. As shown in figures 2 and 4A, the five C repeats located 5' to the ξ1 gene have inserted into two sites that are only 19 bp apart. Comparisons with a consensus sequence for C repeats (see below) and the assignment of flanking direct repeats indicate that C19 inserted into C20, which previously had inserted into C18. Immediately 5' to those repeats, C16 inserted into C17. Thus, three repeats have inserted into two preexisting ones. Figure 3 shows three C repeats located 5' to the ζ2 gene that have
Fig. 1.—Map of rabbit α-like globin gene cluster. Related genes are shown as boxes with the same shading. BamHI sites are marked along the middle line; solid fill along this line indicates that the region has been sequenced. J sequences are drawn as open, pointed boxes. C repeats are shown as solid triangles that point in the direction of their A-rich 3' tract, and the numbers assigned to the sequenced repeats are shown beneath them. A rabbit repetitive element found by hybridization experiments (Cheng et al. 1987) is shown as a stippled polygon between the genes C2 and C3; current data do not exclude the possibility that it is one or several C repeats. The lower line shows the T homology blocks containing 8 genes and the Z homology blocks containing 9 genes, separated by J sequences at the junctions (Cheng et al. 1987).
Fig. 2.—Sequence of region containing five C repeats 5' to rabbit ζ1 globin gene. Flanking direct repeats marking the beginning and end of each C repeat are shown in boldface, with the number of the repeat they flank written above them. Repeated sequences and their flanking direct repeats are in capital letters, whereas nonrepetitive DNA is in lowercase letters. Arrows at the side of the sequence show the position and orientation of repeat sequences; recessive insertions are shown as interruptions in the arrows. Members of the least divergent (and hence proposed to be most recently propagated) subfamily of repeats, group III, are shown as solid arrows, a gray filled arrow represents a group II repeat, and white arrows represent members of the most divergent (and hence proposed to be oldest) subfamily, group I. Arrowheads point toward the A-rich tracts of the repeats. Sequences 3' to position 1965 were previously reported by Cheng et al. (1988) along with the remainder of the ζ1 gene sequence, whose coding region begins at position 2336. The TATA box, cap sites, and ATG initiation codon of the ζ1 gene are underlined.
FIG. 3.—Sequence of region containing three C repeats 5' to rabbit $z^2$ globin gene. Conventions used are the same as those in fig. 2. Sequence 3' to C25 is part of the first intron of the truncated rabbit $z^2$ gene, as reported by Cheng et al. (1988).
subfamilies and clustering of rabbit C repeats

Fig. 4.—Recursive insertion of C repeats in three different locations within rabbit genome. Repeats that are members of the oldest subfamily (group I) are shown as white arrows, members of the next oldest subfamily (group II) are shown as gray arrows, and repeats from the most recently propagated subfamily (group III) are shown as black arrows. Heads of the arrows point in the direction of the repeats' A-rich 3' tracts. A, Diagram of region immediately 5' to rabbit ζ1 gene in α-like globin gene cluster. B, Diagram of cluster of three repeats between ε and γ genes in rabbit β-like globin gene cluster. C, Diagram of five repeats between γ and δ genes in rabbit β-like globin gene cluster.

inserted into a sequence of 200-bp single-copy DNA. Although the 5' end of C23 has not been sequenced, it is clear that C23 and C24 are juxtaposed. The β-like globin gene cluster provides several additional examples of insertions into or in close proximity to other repeats (figs. 4B and 4C). In all cases the repeats which have inserted into preexisting repeats are members of the least divergent—and hence proposed to be more recently propagated—subfamilies of repeats (see below). No instance of a repeat from an older subfamily inserting into a member of a newer subfamily is seen. In several cases, the new C repeat has inserted into the 5' end of the preexisting repeat.

If each C repeat inserted randomly into the genome—i.e., in a manner independent of flanking sequences and regardless of whether another C repeat had previously inserted nearby—the distribution of the number of repeats found within a window of arbitrary size should fit a Poisson distribution. The non-C repeat DNA in the α- and
β-like gene clusters was divided into 50-bp intervals, and the number of repeats within each window was determined (table 2). For a random distribution of repeats, the insertion of five repeats into one 50-bp window of single-copy DNA in the α cluster (14.4 kb total sequence length), as is seen 5' to the ζ gene, is expected to occur $2.48 \times 10^{-9}$ times, and the tail probability of such an event is $<10^{-36}$. The insertion of three repeats into a 50-bp window is expected to occur only $3.47 \times 10^{-3}$ times in the combined data set, yet it has occurred twice (5' to ζ gene and between the δ and γ genes). The distribution of C repeats both in these globin gene clusters and in the uteroglobin gene deviates markedly from a Poisson distribution (table 2). Thus, we conclude that C repeats do not insert randomly but, rather, have a strong tendency to insert where others have inserted before, thereby forming clusters. Other statistical tests of the distribution of C repeats in these regions which consider the fact that the previous insertion of a repeat increases the effective window size for a secondary insertion also strongly support the conclusion that C repeats have a strong tendency to cluster (Krane 1990).

Internal Organization of C Repeats

Previous studies with a small set of C repeats had indicated that they may be subdivided into families (Cheng et al. 1984; Hardison and Printz 1985), and, with the new sequence data from the α- and β-like globin gene clusters, as well as from other sources (table 1), it became possible to examine this issue more rigorously. An alignment of 45 different C repeat sequences is shown in figure 5. The beginning and end of most C repeats are demarcated by associated flanking direct repeats that range in length from four to 19 nucleotides. While nine repeats begin at the first of four G's that start the consensus, the majority of the repeats begin shortly before or after; hence a precise 5' end cannot be defined. A unique 3' end also cannot be assigned, because of variations in the size of the A-rich tracts, ranging from one to 73 residues. Insertions and deletions can be seen at several sites, most notably between positions 80 and 81 and between positions 211 and 212. The average length of the 39 nontruncated rabbit C repeats is 324 nucleotides.

Four subregions of C repeats can be distinguished, some of which are significantly similar to sequences in other organisms. The sequence between positions 7 and 94 in the consensus is related to tRNA genes (fig. 6). This sequence is 66% similar to a human glycine tRNA, corresponding to a region previously identified as similar to tRNA genes (Kazuichi and Okada 1985). While this region has more ambiguous nucleotides than any other part of the C repeat consensus sequence, it does include the conserved boxes A and B that comprise the internal promoter for RNA polymerase III (Ciliberto et al. 1983). Adjacent to this tRNA-like segment is a "core region" (position 95 through position 264; fig. 5) that is well conserved among rabbit C repeats. It is surprising that the "bovine consensus sequence" of the major class of artiodactyl SINE (Spence et al. 1985) is similar to this core region. The bovine consensus sequence is often found associated with a tRNA-like sequence (artiodactyl "C") in bovine genomic DNA (Wanatabe et al. 1982). The tRNA-like and core regions together form a contiguous sequence that is 56.9% similar between these two families of repeats (fig. 6). Although this is a notable match, it is of marginal statistical significance. One hundred random scramblings of the artiodactyl consensus sequence and subsequent alignment to the C repeat consensus resulted in three matches of equal or higher similarity. Hence, while these sequences show some similarity, the alignment does not provide an unequivocal argument for common ancestry.
Table 2
Observed and Expected Occurrences of Insertion of C Repeats into 50-bp Windows

<table>
<thead>
<tr>
<th>No. of Repeats in a 50-bp Window</th>
<th>α Cluster</th>
<th>β Cluster</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>216</td>
<td>209</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>10.5</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.262</td>
<td>0.1654</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4.36 × 10⁻³</td>
<td>0.0017</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>5.45 × 10⁻⁵</td>
<td>...</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2.48 × 10⁻⁹</td>
<td>&lt;10⁻³⁶</td>
</tr>
<tr>
<td>≥6</td>
<td>0</td>
<td>&lt;4.5 × 10⁻⁵</td>
<td>...</td>
</tr>
</tbody>
</table>

* From the 11 repeats that have inserted into the 14.4-kb (11 kb of single-copy DNA) sequenced region.
* If repeats had a Poisson distribution with the observed mean number of repeats/50-bp window used as the Poisson parameter.
* Probability of observing one or more 50-bp window with the given number of repeats, when this same Poisson distribution is assumed.
* From the 15 repeats that have inserted into the 44.6 kb (39.6 kb of non-C repeat DNA) sequenced region.
* Includes the three insertions that have occurred in the rabbit uteroglobin gene.
FIG. 5.
FIG. 5.—Multiple alignment of 45 C repeats and an overall consensus sequence. A period (.) indicates an identical match with the consensus (ConC) nucleotide at that position, while nucleotides are shown for positions that differ from the consensus or are represented as ambiguous positions in ConC (R = G or A; Y = C or T; K = G or T; W = A or T; S = G or C; M = A or C; B = not A; D = not C; V = not T; H = not G; N = G, A, C, or T). Dashes (-) indicate gaps inserted to improve the alignment. Sites diagnostic for subfamilies are underscored in ConC, and the consensus sequences for internal polymerase III promoters (Box A and Box B) (Ciliberto et al. 1983) are shown above ConC. The ConC sequence is numbered above the lines. Lowercase nucleotides at the beginnings and ends of sequences are flanking direct repeats, and the delta symbol (Δ) indicates the end of available sequence data. The double dagger (‡) and asterisk (*) symbols indicate the positions at which the correspondingly labeled sequences at the bottom of the figure were deleted, for clarity of presentation, from the aligned sequences.
Consensus sequences for the internal promoters required for transcription by RNA polymerase III (Box A and Box B) are also shown above the sequences. A solid bar (I) between nucleotides indicates they are identical; a colon (:) indicates a “half-matches” between a unique and a twofold ambiguous nucleotide; and a period (.) indicates a match between a unique and a threefold ambiguous nucleotide or between two different twofold ambiguous positions.
Following the core region is a segment that contains the repeating dinucleotide \((\text{CT})_n\) in all full-length C repeats except C26 (fig. 5). The CT dinucleotide is repeated from 3–51 times. Finally, at the 3’ end of most full-length C repeats is an A-rich segment that includes two tandem and overlapping AATAAA polyadenylation signals in the consensus (position 313 through position 322; fig. 5), although individual repeats may have more.

A few rabbit C repeats appear to have inserted as 5’ truncated sequences. The 5’ flanking direct repeats for C25 and C2 indicate that they begin at positions 102 and 103, respectively (fig. 5). Hence, they are missing the tRNA-like region and begin close to the core regions of the repeat.

Subfamilies of C Repeats

Two different approaches to phylogenetic analysis demonstrate that C repeats can be divided into subfamilies. A cluster analysis (Hartigan 1983) of pairwise percentage identities (table 3) among all full-length C repeats generated the dendrogram shown in figure 7, left. Nine repeats (C12–C23; fig. 7, left) form an obvious group (III) whose sequences are very similar among themselves. Another set of 13 repeats (C31–C6) form group II in the cluster analysis (fig. 7, left), although its members are less similar to one another (table 3). In most cases the remaining 15 C repeats (C11–C9; fig. 7, left) are as different from each other as they are from members of groups II and III (table 3) and are classified as members of group I.

The method of maximum parsimony (Swofford 1990) produced two equally parsimonious trees which differed in only minor detail; one tree is shown in figure 7, right. The consensus sequence of the predominant SINE in artiodactyls was used as an outgroup because of its similarity to the rabbit C repeat consensus. Again, two well-defined clusters of repeats are seen in the two PAUP trees; 13 C repeats are in group II and 10 are in group III. The remaining 19 repeats are assigned to the more heterogeneous group I. Assignments by the parsimony method are the same as those made in the cluster analysis, with the exceptions of C5, which is assigned to group II by clustering but to group I by PAUP, and C35, which could not be unambiguously assigned by maximum parsimony. The horizontal branch lengths are proportional to the number of nucleotide differences between repeats, and the relative branch lengths within groups are also the same as those in figure 7, left. Bootstrapping, which checks the internal consistency of the data, further supports the assignment of all the repeats except C35 to only one of three groups (data not shown). Thus, these two different approaches—cluster analysis and parsimony—give very similar assignments of rabbit C repeats to subfamilies.

While most repeats in group I differ considerably from each other, some pairs are quite similar, including C20/C24 and C17/C36 (fig. 7). Both of these pairs are in a part of the rabbit \(\alpha\)-like globin gene cluster that was generated by block duplications of both genes and repeats (Cheng et al. 1987). Thus, their divergence likely corresponds to the time since a duplication event. Also, the divergence between rabbit C15 and its homologue in hare, H15 (fig. 7, right), reflects the time these species have been separated.

As reflected by the branch lengths in both phylograms in figure 7, the extent of intragroup similarities varies considerably among the three subfamilies of C repeats. The similarities among group I repeats are much lower than those in group II intragroup comparisons, which in turn are lower than the similarities among group III repeats (table 3). The average identity of full-length C repeats to each other, only 59%, is
### Table 3
Matrix of Percent Identities among Full-Length C Repeats

<table>
<thead>
<tr>
<th></th>
<th>Group III</th>
<th>Group II</th>
<th>Group I</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>80</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>84</td>
<td>85</td>
<td>100</td>
</tr>
<tr>
<td>16</td>
<td>71</td>
<td>75</td>
<td>97</td>
</tr>
<tr>
<td>38</td>
<td>74</td>
<td>74</td>
<td>81</td>
</tr>
<tr>
<td>23</td>
<td>72</td>
<td>80</td>
<td>79</td>
</tr>
<tr>
<td>29</td>
<td>72</td>
<td>74</td>
<td>73</td>
</tr>
<tr>
<td>30</td>
<td>68</td>
<td>77</td>
<td>75</td>
</tr>
</tbody>
</table>

**Group III:**
- 12: 100
- 8: 80 100
- 3: 84 85 100
- 16: 73 71 64 100
- 38: 71 75 79 79 100
- 14: 74 74 81 81 84 100
- 23: 72 80 79 78 83 86 100
- 29: 72 74 73 74 91 31 84 100
- 30: 68 77 75 74 85 79 85 92 100

**Group II:**
- 31: 65 68 70 63 54 65 60 59 56 100
- 42: 69 75 78 66 68 72 70 72 65 74 100
- 40: 68 71 75 61 70 67 71 65 63 78 81 100
- 44: 70 72 76 64 65 74 69 63 63 75 78 79 100
- 7: 65 73 77 68 63 65 66 66 61 73 75 77 74 100
- 34: 62 69 68 58 68 67 57 60 60 67 70 78 76 71 100
- 35: 64 69 72 63 67 68 62 65 62 73 76 70 72 72 72 100
- 28: 68 70 74 66 69 66 67 65 65 71 77 76 73 73 75 76 100
- 19: 65 70 72 57 66 61 64 57 56 67 75 72 67 69 71 70 72 100
- 21: 61 58 64 60 60 61 55 64 61 64 71 72 71 66 70 65 73 70 100
- 37: 69 66 71 63 69 67 62 68 62 69 74 74 75 73 74 77 73 71 100
- 22: 60 63 67 56 58 59 60 58 54 69 74 70 68 68 64 63 63 66 62 68 100
- 6: 55 60 61 51 54 61 54 49 52 62 70 67 67 66 68 63 58 66 59 71 60 100
- 5: 61 63 64 50 55 60 58 56 53 59 71 72 71 70 67 67 54 66 63 70 62 65 100
<table>
<thead>
<tr>
<th>Group 1:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>61 62 64 56 53 49 55 53 48 58 69 64 62 62 67 63 64 63 65 55 60 67 100</td>
</tr>
<tr>
<td>43</td>
<td>61 53 56 47 59 58 53 57 54 57 67 64 69 65 63 59 63 64 59 69 49 54 66 61 100</td>
</tr>
<tr>
<td>15</td>
<td>51 55 53 43 47 52 50 44 41 56 61 54 56 58 53 57 57 60 56 57 55 62 60 60 52 100</td>
</tr>
<tr>
<td>32</td>
<td>57 62 60 52 50 56 54 52 45 56 70 57 62 58 56 65 58 63 58 66 49 56 60 61 63 63 100</td>
</tr>
<tr>
<td>10</td>
<td>57 56 55 48 52 52 48 46 48 56 57 57 58 56 57 59 59 62 57 58 57 57 59 59 56 63 100</td>
</tr>
<tr>
<td>26</td>
<td>51 53 57 44 50 53 56 55 53 57 63 56 58 59 57 56 58 58 50 59 53 53 59 53 52 60 61 63 100</td>
</tr>
<tr>
<td>18</td>
<td>55 53 50 51 47 50 47 48 46 44 58 58 61 51 60 57 55 55 57 56 56 52 58 55 55 47 60 47 56 100</td>
</tr>
<tr>
<td>1</td>
<td>45 47 44 46 43 45 45 43 43 41 53 54 48 47 47 47 49 45 47 47 50 49 51 52 54 55 51 53 54 47 51 100</td>
</tr>
<tr>
<td>4</td>
<td>49 51 53 42 40 46 36 43 43 42 49 54 50 53 47 50 45 53 50 51 46 44 61 47 51 59 45 45 45 45 49 54 100</td>
</tr>
<tr>
<td>41</td>
<td>51 58 52 46 46 50 43 53 46 53 61 59 59 58 61 56 52 59 54 48 50 56 56 53 57 58 61 53 53 57 58 60 100</td>
</tr>
<tr>
<td>17</td>
<td>51 53 59 45 56 53 44 52 46 44 57 53 53 54 56 56 54 57 43 54 46 53 59 50 40 55 59 48 50 51 44 47 51 100</td>
</tr>
<tr>
<td>36</td>
<td>47 47 45 42 42 42 40 43 48 50 55 52 52 56 53 56 47 46 49 47 52 56 51 54 48 44 50 54 51 47 50 42 44 50 70 100</td>
</tr>
<tr>
<td>20</td>
<td>46 47 52 49 48 53 46 50 44 58 60 57 49 54 59 63 62 57 58 56 51 47 58 49 49 38 48 51 43 45 45 50 39 45 50 100</td>
</tr>
<tr>
<td>24</td>
<td>54 55 59 54 54 56 49 53 45 62 64 58 60 61 58 58 56 58 59 56 59 54 55 54 54 53 55 49 51 49 45 41 52 45 51 91 100</td>
</tr>
<tr>
<td>9</td>
<td>45 51 48 46 46 47 42 42 38 45 45 49 44 55 54 56 54 53 49 59 44 47 46 48 50 49 58 56 51 44 50 42 52 44 41 44 44 100</td>
</tr>
</tbody>
</table>

**NOTE.**—Both in the column heads and in the first-column entries, the numbers subsumed by the group designations are the numbers of each C repeat.
considerably less than that reported for \textit{Alu} repeats (Schmid and Shen 1985; Hwu et al. 1986; Jurka and Smith 1988) and other SINEs (Deininger 1989). The percentage identities suggest that group I repeats have diverged independently for 43–131 Myr (average 77 Myr), group II repeats for 22–59 Myr (average 37 Myr), and group III repeats for 8–42 Myr (average 26 Myr), assuming a neutral substitution rate of 5
PHYLOGENETIC TREES INDICATING SUBFAMILY RELATIONSHIPS OF RABBIT C REPEATS. HORIZONTAL DISTANCE IS PROPORTIONAL TO THE EVOLUTIONARY DISTANCE SEPARATING REPEATS. SUBFAMILIES TO WHICH REPEATS HAVE BEEN ASSIGNED ARE BRACKETED AND LABELED. LEFT. DENDROGRAM GENERATED BY CLUSTER ANALYSIS OF AVERAGE SIMILARITY AMONG FULL-LENGTH REPEATS. RIGHT. ONE OF TWO EQUALLY PARSIMONIOUS PHYLOGRAMS GENERATED BY PAUP PROGRAM BY USING CONSENSUS FOR PREDOMINANT SINE IN ARTIODACTYLS (ARSIN) AS OUTGROUP. THE OTHER EQUALLY PARSIMONIOUS TREE DIFFERED PRIMARILY IN THE PLACEMENT OF REPEATS C34 AND C42 WITHIN GROUP II (DATA NOT SHOWN). THE ORTHOLOG OF C15 IN HALES IS DESIGNATED H15.
Table 4
Nucleotides in Diagnostic Sites of 44 Rabbit C Repeats

<table>
<thead>
<tr>
<th>NUCLEOTIDE</th>
<th>GROUP III</th>
<th>GROUP</th>
<th>GROUP III</th>
<th>GROUP</th>
<th>GROUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Consensus</td>
<td>8 12 16 20 24 28 32 36 40 44 48 52 56 60</td>
<td>3 7 11 15 19 23 27 31 35 39 43 47 51 55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group III Consensus</td>
<td>2 6 10 14 18 22 26 30 34 38 42 46 50 54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**NOTE.**—In the column heads the numbers subsumed by the group designations are the numbers of repeats. A dot (·) represents a match with the group III consensus nucleotide, and a letter denotes a nonmatching nucleotide. A dash (−) indicates nucleotides missing because of gaps or truncations in the sequences. The first column ("Position") lists three sets of sites: those diagnostic for group III (position 12 through position 309), those diagnostic for group I (position 1 through position 189), and that distinctive for each of the three groups (position 156).

$\times 10^{-9}$ substitutions/site/year (Li et al. 1981). These relative ages suggest a model in which group I repeats were generated prior to group II repeats, which in turn were propagated prior to group III repeats.

Given that group I of the C repeats is operationally defined as all those that do not fit into either group II or group III, it is likely that it consists of several very old subfamilies. However, these additional subfamilies are very difficult to resolve. In fact, some of the substructure seen within group I in the PAUP tree of figure 7, right, is not seen in the cluster tree of figure 7, left. Hence no attempt has been made to subdivide this heterogeneous group.

**Diagnostic Sites**

To confirm the groupings seen in the parsimony and cluster analyses at the nucleotide level, sites distinctive for each group were identified. A total of 25 such positions (table 4) were found that allowed placement of ≥85% of the repeats into the same groups as were defined by the two tree-generating methods. No single site allows definitive assignment of all C repeats to the correct group, but the assemblage of diagnostic
sites does allow unambiguous classification. The power of the diagnostic sites to discriminate between groups was tested as follows: First the consensus sequences at these sites for the three groups were determined, and then the similarity of each C repeat in table 4 to each of the three consensus sequences was scored. In all cases the similarity to the “correct” consensus was considerably greater than the similarity to the other two consensus sequences. Hence the assignment of individual C repeats to a particular group was unambiguous. In a further test, repeats discovered after the diagnostic sites were identified (repeats C37 and C39–C45) were assigned to groups on the basis of these sites alone, and the assignments were subsequently confirmed by both tree-generating programs. The repeats that were difficult to assign to groups on the basis of cluster or parsimony analysis alone, C5 and C35, were placed unambiguously into group I and group II, respectively, by the criterion of similarity at the diagnostic sites.

These 25 diagnostic sites show evidence of a progressive but punctuated change in the three subfamilies of C repeats (table 4). Seventeen sites distinguish groups I and II from group III, seven sites distinguish groups II and III from group I, and one site is unique to each subfamily. For example, at position 30 in the consensus, all repeats assigned to group III have a T, whereas most repeats in groups I and II have a G. Similarly, nucleotides corresponding to position 186 in group I repeats are mostly G, whereas those of groups II and III in the same position are all A. The seven diagnostic sites in common between group III and II indicate that the latter are more closely related to each other than to group I and that these seven sites were changed before the activation of the progenitor to the group II repeats. Likewise, at least another 17 sites were changed in the progenitor of group III relative to the progenitor of group
II. As none of the 44 repeats analyzed represents a clear intermediate between these three groups, the changes in the progenitors appear to have been punctuated (i.e., occurred abruptly relative to the duration of propagation of a group of repeats).

Copy Number of C Repeats

Given the amount of divergence between subfamilies, the copy number of C repeats in the rabbit genome was reexamined by quantitative DNA hybridization experiments. Probes for each subfamily (Cl for group I, C19 for group II, and C2 for group III) were hybridized to titrations of genomic DNA and cloned DNA from repeats representing each subfamily. Quantitative analysis of the autoradiograms shows that repeats hybridizing to C1 (group I) constitute 0.27% of the rabbit genome (26,000 copies), those hybridizing to C19 (group II) constitute 0.43% of the genome (38,000 copies), and those hybridizing to C2 (group III) constitute 1.84% (170,000 copies) of the genome (Krane 1990). The probes for different subfamilies did not cross-hybridize, so the copy numbers of all three of the subfamilies can be added to give a minimal estimate of at least 234,000 copies of C repeats/haploid genome.

However, the probes used for group I and group II C repeats are probably not hybridizing to a major portion of the members of those subfamilies because of the increased diversity seen in older subfamilies. The contribution of each subfamily to the total number of C repeats in the genome can be estimated from the phylogenetic analysis shown in figure 7. Of these C repeats, 41% are classified as members of group I, 30% as group II, and 18% as group III. If these numbers apply to the total population of C repeats, then one can extrapolate from the 170,000 copies of the least divergent group III that the entire C repeat family has ~1 million repeats/haploid genome.

C repeats constitute 14.3% (9,390 of 65,611 nucleotides) of three regions of long contiguous sequence available from the rabbit genome (Suske et al. 1983; Margot et al. 1989; Krane 1990). This 14.3% of the haploid genome of rabbits is equivalent to 1,300,000 copies of C repeats, which is in good agreement with the extrapolation from the hybridization experiments.

Discussion

A multiple alignment of 44 members of the rabbit C repeat family of SINEs has revealed much information on the origin and propagation of these elements. The C repeats are composites of sequences that show four distinguishable regions, some of which are related to sequences in other species. At least three distinct subfamilies have been identified, indicating that a limited number of progenitors have propagated at different times to form the most divergent subfamily (group I) first, followed by another round of retrotransposition to form group II and, finally, the most recent propagation, that of the group III subfamily. The relative ages of the subfamilies are confirmed by analysis of recursive integrations of C repeats—in no case has a repeat assigned to an older subfamily inserted into a member of a younger subfamily.

The high similarity between the new C repeat consensus sequence and tRNA genes supports the proposal that the 5' end of C repeats may be derived from a tRNA gene (Kazuichi and Okada 1985). A conserved RNA polymerase III internal promoter may direct the transcription of some repeats, but two repeats (C2 and C25) inserted into the rabbit genome do not contain this promoter, indicating that it is not required for the actual insertion of the repeat. Searches of GenBank have shown that the core region of C repeats is similar to a portion of a highly repeated sequence in artiodactyls.
This region of the repeats may play a common role in retrotransposition in both lagomorphs and artiodactyls.

Another region of the repeats, the A-rich tract, may be added posttranscriptionally. Conserved polyadenylation signals at the 3' end of the C repeat sequence are occasionally used for the polyadenylation of messages of genes into which the repeats have inserted (Krane and Hardison 1990); hence they may be a signal for the polyadenylation of C repeat transcripts as well. Similar polyadenylation sites are also found in B1 repeats and B2 repeats (Ryskov et al. 1984), indicating that polyadenylation may be important in the retransposition of several families of SINES. Posttranscriptional cleavage and polyadenylation directed by the AAUAAA signal would preclude intramolecular priming by hybridizing the A-rich tract to the string of three or four U's at the 3' end of the transcript that result from termination by RNA polymerase III. This intramolecular priming is part of a common model for the reverse transcription of these repeats (Jagadeeswaran et al. 1981). However, primers for reverse transcription could still be provided either by the genomic DNA at staggered breaks (Moos and Gallwitz 1983) or by a simple sequence synthesized de novo at the target site, a situation similar to that seen at telomeres (Rogers 1985).

Calculations based on estimated neutral drift rates indicate that the oldest of three C repeat subfamilies, group I, has been propagated since close to or even before the mammalian radiation, ~85 Mya (Romero-Herrera et al. 1973). However, no C repeat sequences have been found in other mammalian orders (Cheng et al. 1984; Hardison and Printz 1985), and data-base searches with the improved consensus show only strong matches with rabbit and hare sequences. Thus, it is likely that C repeats began to appear in high copy number in lagomorphs after the latter last shared a common ancestor with other mammals. Once inserted into the genome, the repeats have diverged at the same rate as nonselected regions, as is seen with C15 and Hare 15 (Krane 1990); a similar result has been observed for Alu repeats in primates (Sawada et al. 1985). The rates of sequence divergence differ from one locus to another (Grula et al. 1982; Bohr et al. 1985), and the estimates of two C repeats being separated for >85 Myr may result from the repeats inserting into different regions and/or evolving at a faster rate, closer to the higher estimate of $13 \times 10^{-9}$ substitutions/site/year for pseudogene divergence (Miyata and Yasunaga 1981).

The amount of divergence among the C repeats studied here is greater than that seen for other families of mammalian SINES. The average pairwise mismatch for 39 rabbit C repeats is 41%, ranging from an average of 22% for group III to 48% for group I. In contrast, the pairwise mismatch values for human Alu repeats range from 1.8% to 34.8% (Slagel et al. 1987; Deininger and Slagel 1988). The average divergence from the consensus is 16% for human Alu repeats (Deininger 1989) and 10% for rodent B2 repeats (Rogers 1985). The rabbit C repeats were sequenced because of their association with genes or gene clusters; hence they should not be biased for any subfamily, as would repeats selected by hybridization. The accumulating data bases on primate and rodent SINES are now also predominantly from sequenced regions and should be comparable to that for rabbit C repeats. The greater divergence of rabbit C repeat sequences relative to SINE families in other mammalian orders therefore indicates that the C repeats may be older.

Both rabbit C repeats (Margot et al. 1989; present paper) and human Alu repeats (Slagel et al. 1987; Stoppa-Lyonnet et al. 1990) show a pronounced tendency to insert into regions where other repeats have inserted before. The clustering of C repeats may result from DNA structure alterations brought about by the repeats' characteristic
(CT)$_n$, stretches and A-rich tracts. Tandemly repeating homopurine/homopyrimidine tracts such as (CT)$_n$ can fold into the triple-helical H-form DNA (Htun and Dahlberg 1988), and poly A-tracts tend to be nucleosome free (Kunkel and Martinson 1981). These disruptions in DNA or chromatin structure could mark a preferred region for insertion of new repeats. If so, this appears to act over a few hundred base pairs, since several C repeats insert into the 5' end—not the 3' end—of preexisting repeats. This contrasts with the clustering of human Alu repeats, in which the insertions are usually into A-rich tracts at the 3' ends of other Alu's (Slagel et al. 1987). Another factor that could contribute to the clustering is a selection against insertions in the regions that do not contain C repeats.

A regional preference for SINES to insert into G+C-rich regions and for LINEs to insert into A+T-rich regions has been noted in cytological studies in humans (Korenberg and Rykowski 1988), and a similar bias is also seen in the sequences of the rabbit α- and β-like globin gene clusters. C repeats contribute 24% of the sequenced region of the G+C-rich rabbit α-like gene cluster—but only 10% of the A+T-rich β-like gene cluster. In addition, no sequence similar to the predominant LINE in rabbits has been found by either sequence analysis or hybridization within the rabbit α-like gene cluster (Cheng et al. 1987), while portions of six have inserted into the β-like gene cluster (Margot et al. 1989; Huang et al., accepted). Mammalian and avian α-like globin gene clusters are found in G+C-rich isochores (Bernardi et al. 1985). These isochores are regions of compositional homogeneity over very long (≥200-kb) stretches of DNA, and it is likely that their formation predates the propagation of SINES in warm-blooded vertebrates (Bernardi et al. 1985; Bernardi 1989). However, a bias for insertion of these G+C-rich repeats into already G+C-rich regions could play a role in the maintenance of these dense isochores.

The presence of subfamilies with diagnostic sites that change progressively suggests a model for the propagation of these repeats throughout the rabbit genome. The grouping into subfamilies indicates that only a small number of progenitors produced most members of the C repeat family. The alternative model, with a large number of source sequences, predicts a broad distribution of repeated sequences that could not be resolved into a small number of subfamilies (Kaplan and Hudson 1989). Thus, a single progenitor (or a small number of them undergoing concerted evolution) has produced RNA for retrotransposition, while new copies that have inserted into the genome are usually not capable of producing new repeats—and thereafter accumulate random mutations in the genome. Eventually these mutations may “reactivate” a copied repeat so it becomes a transcriptionally and transpositionally active source gene, producing a new subfamily of C repeats. This would generate the punctuated and progressive changes seen in the diagnostic positions (table 4).

The SINES in different mammalian orders have different origins (Daniels and Deininger 1985; Schmid and Shen 1985) and have propagated independently (Margot et al. 1989), but, as shown here, the rabbit C repeats parallel SINES in other orders, not only in their general structure and mechanism of transposition but also in their evolutionary history (progressive waves of dispersion) and copy number (~1 million). Further exploration of how these independent transposable elements have maintained these parallels in their evolutionary history may reveal information about the roles they play in the genome.
Acknowledgments

We are grateful to Danielle Pasqualone and Douglas Price for the use of their unpublished data and to David Swofford for the use of his PAUP program. This work was supported by Public Health Service grant DK27635 and by RCDA grants HD00743 to A.G.C. and DK01589 to R.C.H.

LITERATURE CITED


MOOS, M., and D. GALLWITZ. 1983. Structure of two human β-actin related processed genes one of which is located next to a simple repetitive sequence. EMBO J. 2:757–761.


ALAN M. WEINER, reviewing editor

Received April 3, 1990; revision received August 16, 1990

Accepted August 16, 1990
Comparison of Alcohol Dehydrogenase Expression in
*Drosophila melanogaster* and *D. simulans*¹

*M. Scott Thomson,² James W. Jacobson,³ and Cathy C. Laurie*
Department of Zoology, Duke University

Alcohol dehydrogenase (ADH) gene expression was analyzed in *Drosophila melanogaster* and its sibling species *D. simulans*. The levels of ADH activity, ADH-cross-reacting material (CRM), and ADH-mRNA were analyzed for several strains of each species, which derive from diverse geographic locations around the world. There is considerable quantitative variation in ADH activity, CRM level, and RNA level among strains within species at all developmental stages. However, the only consistent differences between the two species are in pupal RNA level and in late-adult activity and CRM level. Late-adult *melanogaster* flies that are homozygous for the Slow allozyme have approximately twice the level of ADH activity and CRM as do *simulans* flies. The regression of activity on CRM over strains is highly significant and essentially the same for each species, which means that most, if not all, of the activity difference between the species is due to a difference in concentration of the ADH protein. In contrast, there is no significant regression of CRM level on mRNA level in adults of either species; nor is there a significant difference in RNA level between species. Therefore, the difference in ADH protein concentration is not due to RNA template availability. Thus, the interspecific difference in ADH level in adults must be due either to a difference in the rate of translation of the two RNAs or to a difference in protein stability.

**Introduction**

*Drosophila melanogaster* and its sibling species *D. simulans* are essentially cosmopolitan in distribution and live in close association with human populations (Patterson and Stone 1952). Both species utilize fermenting fruits in which ethanol concentrations range up to several percent by volume (McKechnie and Morgan 1982; David and Van Herrewege 1983). In laboratory tests, *melanogaster* adults consistently show a higher level of tolerance both to ethanol and to other alcohols than do *simulans* adults (McKenzie and Parsons 1972; David et al. 1974; David and Bocquet 1976; Daggard 1981; Gibson and Wilks 1988). Tolerance in larvae has been measured for only a small number of strains, but in those cases *melanogaster* also shows a higher tolerance to ethanol than does *simulans* (McKenzie and Parsons 1972; Parsons et al. 1979; Parsons 1980). Several authors have suggested that the difference in alcohol tolerance detected in the laboratory may affect the distribution of the two species in nature. A consistent observation in studies of winery populations of *Drosophila* is that

1. Key words: alcohol dehydrogenase, *Drosophila melanogaster*, *D. simulans*, gene regulation, interspecific variation.

2. Present address: USDA, U.S. Grain Marketing Research Laboratory.

3. Present address: Department of Biology, University of Houston.

Address for correspondence: Cathy C. Laurie, Department of Zoology, Duke University, Durham, North Carolina 27706.

© 1991 by The University of Chicago. All rights reserved.
0737-4038/91/0801-0007$07.00
*melanogaster* is very abundant inside wineries where ethanol concentrations are often very high, while *simulans* is very rare inside compared with immediately outside these same wineries (e.g., see McKenzie and Parsons 1972; McKenzie 1974; Marks et al. 1980; Gibson et al. 1981; Gibson and Wilks 1988). Two additional studies suggest that *melanogaster* larvae are found more frequently than *simulans* larvae in fermenting fruits with high alcohol contents (McKenzie and McKechnie 1979; Oakeshott et al. 1982). It is not clear whether environmental alcohol concentrations actually determine differences in the distributions of these two species on different substrates, but studies of many populations in different parts of the world clearly show that the adults have different tolerances to alcohol under laboratory conditions (see discussions in Gibson et al. 1981; Gibson and Wilks 1988).

The alcohol dehydrogenase enzyme (ADH; E.C. 1.1.1.1) of *D. melanogaster* clearly plays an important role in alcohol detoxification and metabolism. Flies homozygous for an ADH gene (Adh) null allele are extremely sensitive to the toxic effects of environmental alcohols compared with flies possessing an active enzyme (David et al. 1976). More than 90% of the ethanol that is metabolized to lipid in larvae goes through a pathway that is dependent on ADH activity (Geer et al. 1985). Furthermore, genetic variation in ADH activity levels in *melanogaster*, particularly the difference between allozymes, is frequently associated with variation in tolerance both to ethanol and other alcohols (see reviews in Gibson and Oakeshott 1982; Van Delden 1982).

Several investigators have reported that *melanogaster* adults have much higher ADH activity than do *simulans* adults (Pipkin and Hewitt 1972; McDonald and Avise 1976; Juan and Gonzalez-Duarte 1980; Daggard 1981; Oakeshott et al. 1982; Gelfand and McDonald 1983; Chambers et al. 1984b; Dickinson et al. 1984; Gibson and Wilks 1988). From these studies it appears that Adh*F* *melanogaster* adults have about twice the *simulans* level of activity and that Adh*S* *melanogaster* adults have about four times the *simulans* level of activity. Many of the authors cited above have suggested that the higher alcohol tolerance of *melanogaster* adults is due to their higher ADH activity.

Because of its potential adaptive significance and because interspecific gene transfer between *melanogaster* and *simulans* is feasible by P-element transformation (Spradling and Rubin 1982; Scavarda and Hartl 1984), we have initiated a study of the molecular basis of the difference in ADH expression between these species. The present report provides a description of the interspecific differences in ADH expression throughout development, at the levels of ADH activity, ADH-protein concentration, and ADH-mRNA concentration.

The *melanogaster* Adh produces two different transcripts, which are differentially expressed during development but produce identical proteins (Benyajati et al. 1983b; Savakis et al. 1986). The sequence of the *Adh* from *simulans* (Bodmer and Ashburner 1984) indicates that it has the same transcriptional organization as *melanogaster* in terms of the positions of the distal and proximal promoters and of the three introns. Dickinson et al. (1984) showed that *simulans* produces primarily the proximal transcript in larvae and primarily the distal transcript in adults, a result which is consistent with the pattern shown by *melanogaster*. In the present report we provide a more detailed comparison of the relative abundances of the distal and proximal transcripts throughout development.

Gel-electrophoresis studies of populations from around the world have revealed no Adh allozyme polymorphism in *simulans*, whereas most *melanogaster* populations are polymorphic for two common allozymes, Adh*F* and Adh*S* (e.g., see Choudhary and Singh 1987). In comparison of ADH expression in *melanogaster* and *simulans*,

**Thomson et al.**
it is important to take into account the major systemic difference in ADH activity and in protein level due to the melanogaster allozyme polymorphism (see Laurie and Stam 1988). The DNA sequence data of Kreitman (1983), Bodmer and Ashburner (1984), and Cohn and Moore (1988) show that the Adh of simulans is more similar to melanogaster AdhS than to AdhF alleles. Among the 18 sites at which the consensus AdhS and AdhF sequences differ, there are 13 matches between simulans and AdhS. One of these matches includes the S/F amino acid substitution site so that the simulans ADH differs from the Slow melanogaster ADH by two amino acids and from the Fast melanogaster ADH by three. These data and the lower levels of polymorphism within AdhF sequences compared with AdhS sequences clearly suggest a recent origin of the high-activity AdhF allele within melanogaster (Kreitman 1983; Aquadro et al. 1986). Therefore, the most direct approach to studying interspecific differences in ADH expression is to compare simulans alleles and melanogaster Slow alleles. Thus, all of the melanogaster lines analyzed in the present report are AdhS.

Material and Methods

Fly Stocks

Seven melanogaster lines were made isoallelic for the Adh region by extracting the entire second chromosome by using a balancer stock; these lines are Was (from Seattle), Jas (from Ishigaki, Japan), Afs (from Burundi, Africa), and Frs (from Bully, France), which were provided by Kreitman (1983), and KA27 (from Lawrence, Kans.), W109 (from Cochrane, Wisc.) and RI42 (from Providence, R.I.), which have been described by Laurie-Ahlberg et al. (1980). Eleven simulans lines were made isoallelic for the Adh region by a procedure that makes use of the close linkage (1.5 map units in melanogaster; Lindsley and Grell 1968) between Adh and the black locus. A b* stock of simulans was obtained from E. H. Grell; this mutant is semidominant. A stock isoallelic for the b region is made by crossing wild-type flies to b*/b* flies, backcrossing a single b*/F1 male to b*/b2 females, intercrossing the b2/+ F2 progeny, and finally intercrossing the +/+ progeny. The wild-type flies for constructing the isoallelic lines were obtained from 11 different isofemale lines: Tun-6 and Tun-16 (Tunisia), Fra-2 and Fra-20 (southern France), Afr S-5 and Afr S-10 (South Africa), and Con S-2 (Congo) were collected in 1983 and provided by R. Singh; Jap-1 and Jap-2 were collected in Munakata City, Japan, in 1986 and provided by T. Yamazaki; Aus-1 was collected in Palmers Island, Australia, in 1986 and provided by J.S.F. Barker; and Ral 1-2 was collected in Raleigh, N.C. in 1984.

Four melanogaster lines homozygous for the Slow ADH allozyme were established by pair-mating flies from four different isofemale lines. After producing progeny, each pair was analyzed by starch-gel electrophoresis (Laurie-Ahlberg and Weir 1979) to ensure that both members were Slow homozygotes. The four melanogaster pair-mated lines were derived from the following isofemale lines: V.France 16-2 (from Vienville, France) and B.W.Africa 27, collected in 1978 and provided by R. Singh; Japan Kochi 5, collected by T. Mukai in 1982; and Ral 124, collected in Raleigh, N.C., in 1982). Four pair-mated simulans lines were also established from the following isofemale lines: S.France 2 (from S.France) and B.Congo S2 (from Brazzaville, Congo), collected in 1983 and provided by R. Singh; Japan Muna-1, collected in 1986 by T. Yamazaki; and Ral S-5, collected in Raleigh, N.C. in 1984. Note that the four pair-mated simulans lines derive from four of the same isofemale lines from which isoallelic stocks were also constructed.
The mutant stock, Adh\textsuperscript{fn23} pr cn, was obtained from W. Sofer. Hochi-R is a highly inbred Adh\textsuperscript{F} strain obtained from W. Doane.

Experimental Design

Experiment 1

One isoallelic line of each species was analyzed at each of 10 developmental stages. The lines were Was (melanogaster) and Ral 1-2 (simulans). For each line and developmental stage a total of four replicate samples were obtained. All samples were obtained as pairs, one from each line, and were reared simultaneously. All 40 pairs of samples are independent in the sense that each was obtained at a different time and from a different culture bottle. Each sample was divided into two parts. One part was homogenized and used for protein assays [ADH activity, cross-reacting material (CRM) level, and total protein level], and the other part was homogenized and used for RNase protection assays (quantitative level of ADH-mRNA relative to an internal standard, Jn23, and visual assessment of the relative abundance of distal and proximal transcripts). Paired t-tests were performed to test the significance of the species difference at each stage. In addition, early- and late-adult data were analyzed together in an analysis-of-variance (ANOVA) model with sources: block (rearing time), sex, age, strain, and their interactions. In the devising of F-tests, all effects were considered fixed except block.

Experiment 2

Four isoallelic lines of melanogaster and 11 isoallelic lines of simulans were sampled as third-instar larvae and adult males. During each of two blocks of time, two replicate samples were obtained from each of the 15 lines. All larval and adult samples were analyzed for ADH activity, CRM level, and total protein level. In addition, RNase protection assays for total ADH-mRNA level were performed on the adult samples from the four melanogaster lines and four of the simulans lines (Cons-2, Jap-1, Fra-2, and Ral 1-2). These eight lines were analyzed to test for species differences. The ANOVA model contained the following main effects and all of their interactions: block, replicate within block, species, and line within species. In the devising of the F-tests, all effects were considered random except species. Eight of the simulans lines, two from each of four geographic locations (France, Japan, Tunisia, and South Africa), were analyzed to test for differences among geographic locations. The ANOVA model contained the following main effects and all of their interactions: block, replicate within block, location, and line within location. In the devising of the F-tests, all effects were considered random. Data from the two stages were analyzed separately.

Experiment 3

Four pair-mated lines from each species, as well as three isoallelic lines (Was, KA27, and Ral 1-2), were sampled as mid-third-instar larvae, mid-pupae, and late adults. The pair-mated lines represent one sample of each species from each of four continents (Europe, Africa, Japan, and North America). Collections of each stage were made independently of one another. During each of two blocks of time, two replicate samples were obtained from each of the 11 lines. Each sample was divided into two parts. One part was homogenized and used for protein assays (ADH activity, CRM level, and total protein level), and the other part was homogenized and used for the mRNase protection assay for total ADH-mRNA level. Data from each of the three stages were analyzed separately. The ANOVA model contained the following
Alcohol Dehydrogenase Expression in Drosophila

main effects and all of their interactions: block, replicate within block, continent, and species. In the devising of the $F$-tests, all effects were considered random except species.

Developmental Staging

All flies were reared on standard cornmeal molasses medium at 25°C. A preliminary experiment was done to determine whether *melanogaster* and *simulans* develop at the same rate. The results are summarized in figure 1 in terms of range and median times for various developmental events. To investigate hatching time, embryos collected over a 1-h period were observed at hourly intervals beginning at 17 h postoviposition. From the curve of number hatched versus time, the median hatch time was estimated. Similar methods were used to estimate median times for the larval moults, pupariation, the appearance of red eye pigment, and eclosion. The results in figure 1 indicate that there are no clear differences in developmental timing between the species except for time of hatching. *Drosophila melanogaster* embryos hatched somewhat later than *D. simulans* embryos, so *Sac* embryos were aged slightly longer than *Ral* 1-2 embryos, in order to sample the same developmental time.

In experiment 2, larvae were not precisely aged but were picked as actively feeding third instars directly from rearing vials. In experiments 1 and 3, staging was accomplished by collecting cohorts of individuals aged at standard intervals past developmental phenomena. The 10 developmental stages of experiment 1 are as follows: *Sac* embryos, 15–19 h postoviposition; *Ral* 1-2 embryos, 14–18 h postoviposition; first instar, 0–4 h posthatch; second instar, 30–32 h posthatch; early third instar, 54–56 h posthatch; mid-third instar, 70–72 h posthatch; prepupae, 0–4 h post–white prepupa; early pupae, 20–24 h post–white prepupa; late pupae, 72–76 h post–white prepupa; early-adult males and females, 0–1 d posteclosion; and late-adult males and females, 7–10 d posteclosion. The three stages of experiment 3 are as follows: mid-third-instar larvae, 70–72 h posthatch; midpupae, 46–66 h post–white prepupa; and late-adult males, 7–10 d posteclosion.

Protein Assays

Samples were prepared for protein assays by grinding freshly collected flies in ice-cold potassium phosphate buffer (10 mM, pH 7.4, 1 mM ethylene diaminetetra-acetate). Homogenates were centrifuged at 15,000 g for 10 min, and the supernatants were stored at $-70^\circ$C.

For assaying ADH activity the spectrophotometric method described by Maroni (1978) was used with isopropanol as substrate. ADH units are expressed as nanomoles NAD$^+$ reduced per minute per milligram total protein. Total protein was determined by the Folin phenol procedure (Lowry et al. 1951).

ADH-protein level was estimated as CRM by radial immunodiffusion (Mancini et al. 1965) using polyclonal rabbit antibodies. A dilution series of a standard fly extract (*Hochi-R, Adh*⁵) was included on each immunodiffusion plate to ensure linearity over the range of sample values. ADH CRM units are given in terms of *Hochi-R* fly equivalents per milligram total protein. The radial immunodiffusion procedure was tested with purified ADH⁵ from *melanogaster* and with purified ADH from *simulans* to verify that there is no difference between the two proteins in extent of antigen-antibody reaction. ADH purification steps were ammonium sulfate precipitation, gel filtration, anion-exchange chromatography, and ATP-affinity chromatography [according to the methods of Lee (1982) and Chambers et al. (1984a)].
RNase Protection Assays

The RNase protection assay to quantitate total ADH-mRNA makes use of the null mutant $Adh^{fn23}$ as an internal control and have been described in detail by Laurie and Stam (1988). The $fn23$ mutant has in exon 3 a 34-bp deletion which causes it
Alcohol Dehydrogenase Expression in Drosophila

to make a defective protein, but it has nearly normal levels of ADH-mRNA (Benyajati et al. 1983a). An RNase protection assay distinguishes between RNAs produced by the deletion-mutant and wild-type alleles. The pBSBD probe extends from the exon 2 BamHI site to an exon 3 Ddel site that lies just 4 bp beyond the 3' end of the fn23 deletion. RNA from wild-type flies protects two regions of this probe: 89 bases of exon 2 and 263 bases of exon 3. When RNA from fn23 homozygotes is used, the protected region from exon 3 is only 225 bases. In the experiments reported here, we used fn23 as an internal control for quantitating total RNA by mixing a homogenate of homozygous adult fn23 males with a homogenate of wild-type individuals of some developmental stage of either simulans or melanogaster. RNA was then isolated from the mixed homogenate, hybridized to radiolabeled pBSBD probe, and digested with RNase, and the protected fragments were separated on an acrylamide gel. Each sample produced both the 263-base fragment derived from the wild type and the 225-base fragment derived from the mutant. The amount of material in each fragment was estimated by cutting bands out of the gel and scintillation counting. Each sample was run as a dilution series, and the amount of wild-type RNA relative to the amount of mutant control was estimated as the slope of the linear regression of number of counts in the wild-type band on number of counts in the mutant band. RNA transcription, hybridization, digestion, and acrylamide-gel analysis were performed by the method of Melton et al. (1984), with modifications described by Laurie and Stam (1988).

Another RNase protection assay utilizes a different probe, pBSHB, which distinguishes proximal and distal ADH messages. The procedure was performed essentially as described by Fischer and Maniatis (1986). The pBSHB riboprobe extends from the HindIII site within the adult intron to the BamHI site in exon 2 (nucleotides 652–1257; Kreitman 1983). This probe protects a 315-base fragment of the second exon from both proximal and distal transcripts, a 134-base fragment specific to the distal transcript, and a 169-base fragment specific to the proximal transcript. The probe is transcribed from the pBSHB plasmid, which was constructed by isolating the 605-base HindIII/BamHI fragment from the pSPHB plasmid [described as pSP6-MEL by Fischer and Maniatis (1986) and provided by J. Posakony] and inserting it into HindIII/BamHI-cut pBSM13 vector (Stratagene, Inc.). RNA transcription, hybridization, digestion, and acrylamide-gel analysis were performed according to a method described by Laurie and Stam (1988), with the following exceptions: The plasmid was linearized with HindIII and was transcribed with T7 polymerase (1.25 units/μl), and hybrid digestions included RNase A (4 μg/ml). These samples were not run in a dilution series, and the bands were not quantitated by counting.

RNA was prepared essentially according to a method described by Fischer and Maniatis (1985), except that no Proteinase K was used. In the following, homogenate refers to the supernatant obtained after first grinding individuals in buffer in a Dounce glass homogenizer and then centrifuging at 12,000 g for 5 min. For experiments 1 and 3, aliquots of a single homogenate of fn23 adults were mixed with each of the wild-type homogenates that were prepared simultaneously (see Experimental Design). For experiment 2, 25 fn23 flies were mixed with 25 wild-type flies prior to homogenization. In experiment 1, the wild-type homogenate was split into two parts; RNA isolated from one part was used for the RNase protection assay to distinguish distal and proximal transcripts, and the other part was mixed with fn23 homogenate and was used for the RNase protection assay to quantitate total message. In the preparation of fn23 and wild-type mixed homogenates for the quantitative assay, the number of fn23 adults per milliliter homogenate was fixed, and the number of wild-type indi-
Results

Experiment 1: Developmental Profile

The developmental profiles of Was (*melanogaster*) and Ral 1-2 (*simulans*) are shown in figure 2. The Was profile for ADH activity is very similar to the *melanogaster* profiles reported previously by Ursprung et al. (1970) for an Oregon-R strain and by Maroni and Stamey (1983) for a Samarkand strain. Activity per milligram total protein rises abruptly during larval development, until the third-instar stage, during which it declines somewhat. Activity falls off rapidly during the pupal stage and begins to rise.
again at eclosion. The Was strain shows a large and significant increase in adult activity between the time points 0–1 d and 7–10 d posteclosion. Figure 2 shows only the average value for adult males and females, but the changes in the two sexes are very similar. Both Ursprung et al. (1970) and Maroni and Stamey (1983) also observed a steep rise in activity after eclosion.

The Ral 1-2 strain shows an ADH activity profile very similar in shape to that of Was, except that it does not show the increase from early to late adult. The difference in activity between early and late adults is not significant. ANOVA of the adult activities for both strains shows that there is a highly significant ($P < 0.001$) age $\times$ strain interaction.

Even though the two strains show developmental profiles of similar shape between embryo and eclosion, there are significant quantitative differences at most of the stages (fig. 2). The direction of the difference in activity between the strains changes twice during development. Compared with Ral 1-2, Was is significantly higher in embryos, consistently lower throughout larval and pupal development, not significantly different at eclosion, and again considerably higher in late adults.

For the two strains the developmental profiles for CRM level are virtually identical to those for activity. CRM level increases significantly ($P < 0.01$) from early to late adulthood in Was and decreases significantly ($P < 0.001$) in Ral 1-2 adults. The RNA profiles are also similar, except that changes are more abrupt and appear to anticipate changes in both activity and CRM, as expected. Furthermore, the rises in both activity and CRM that occur in Was between early and late adulthood are not accompanied by a rise in mRNA. Instead, in both strains, RNA levels show a significant ($P < 0.01$) decline between the early- and late-adult samples. Therefore, the late-adult strain differences in both ADH activity and CRM are not accounted for by a difference in RNA level. In fact, the adult-RNA-level difference goes in the opposite direction. In contrast, the embryonic, larval, and pupal strain differences in both activity and CRM are accounted for by significant RNA-level differences that go in the same direction.

Figure 3 shows that the relative abundance of the proximal and distal transcripts is essentially the same in both strains throughout development. The pattern of relative abundance is very similar to that described by Savakis et al. (1986) for a melanogaster strain. During the embryonic stage the proximal transcript is most abundant, but a low level of distal transcript is also observed. Throughout larval development, until mid third instar, the distal transcript is virtually undetectable, and the proximal transcript is abundant. At the mid-third-instar stage the distal transcript reappears and throughout the pupal and adult stages is clearly the most abundant of the two transcripts. On the basis of the analysis of this one simulans strain, there are no apparent differences between the species in the developmental timing of usage of the distal and proximal Adh promoters.

Experiment 2: Variation among Isoallelic Lines

In experiment 2 four isoallelic lines of melanogaster (all Adh$^S$) and 11 isoallelic lines of simulans were sampled as third-instar larvae and late adults. Figure 4 shows that there is essentially no overlap between the species, in either ADH activity or CRM level, at either developmental stage. The direction of the difference switches between larvae and adults, as observed previously for Was and Ral 1-2. In larvae, the simulans lines have significantly higher activity ($P < 0.001$) and CRM ($P < .05$) levels than do the melanogaster lines. In adults the melanogaster lines are significantly higher for activity ($P < 0.001$) and CRM ($P < 0.05$), but in adult RNA level there
FIG. 3.—Comparison of relative abundances of distal and proximal transcripts during development in Was (melanogaster) and Ral 1-2 (simulans)
FIG. 4.—Variation in ADH expression among isoallelic lines. Circles denote values for *melanogaster* lines (average larvae activity 139; average adult activity 471), and black dots denote values for *simulans* lines (average larvae activity 208; average adult activity 243). The lines represent significant regressions of activity on CRM level over all points, ignoring species. For each regression line, the slope is significantly greater than zero with $P = 0.0001$. $R^2$ values from the regression analyses are given on each plot.

is no significant difference between species. The variation among lines within species is highly significant ($P < 0.01$ in adults; $P < 0.001$ in larvae) for activity at both stages, is marginally significant ($P < 0.06$) for CRM at both stages, and is highly significant ($P < 0.001$) for RNA in adults.

In this experiment the differences detected between the *melanogaster* and *simulans* lines are not necessarily true species differences, since the lines within each species have some genetic background in common, because of the extraction procedure that made them isoallelic. The differences between the two groups of lines might be due to modifier genes contributed by the balancer stock used to isogenize the second chro-
mosomes of the four *melanogaster* lines and by the *b²* stock used to isogenize the *Adh* region of the 11 *simulans* lines.

**Experiment 3: Variation among Pair-mated Lines**

Experiment 3 analyzes variation among four lines of each species that have independent genetic backgrounds. These lines were established from single pair matings of flies from isofemale lines to ensure that each *melanogaster* line is homozygous for the ADH Slow allozyme. The isofemale lines were collected on four different continents. In addition, the isoallelic lines Ral 1-2 (*simulans*), Was, and KA27 (*melanogaster*) were included for comparison with previous results and to determine whether, with respect to ADH expression, they can be considered representative of their species. Table 1 and figure 5 show the lines' means. The results concerning adults are very similar to those described above for the isoallelic lines. Adults from the *melanogaster* pair-mated lines have a significantly higher ADH activity (*P* < 0.01) and CRM level (*P* < 0.01) than do the *simulans* lines, but in RNA level there is no significant difference between species. In contrast, there is no significant difference between species, in either activity or CRM, at either the larval or pupal stage, and figure 5 shows considerable overlap between the species at these stages. Larvae also show no species difference in RNA level, but there is a highly significant difference between species in RNA level at the pupal stage (*P* < 0.0001). Thus we conclude that the larval ADH-level difference observed in experiment 2 is not a true species difference. The only differences that appear to be species specific are activity and CRM in late adults and RNA level in pupae.

In this experiment the species × geographic location interaction in the ANOVA represents variation among lines within a species. This term is significant for larval activity (*P* < 0.0001), larval CRM (*P* < 0.0001), larval RNA (*P* < 0.01), pupal activity (*P* < 0.001), pupal CRM (*P* < 0.0001), and adult RNA (*P* < 0.05). Therefore, most of the ADH expression variables show genetic variation within a species.

Table 1 shows that for each of the variables the isoallelic lines Ral 1-2 and KA27 have values similar to the mean for their respective species and that therefore they

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Mean Activity, CRM, and RNA Levels from Experiment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LINE</strong></td>
<td><strong>LARVAE</strong></td>
</tr>
<tr>
<td><strong>melanogaster:</strong></td>
<td></td>
</tr>
<tr>
<td>V.France 16-2</td>
<td>368</td>
</tr>
<tr>
<td>Japan Kochi 5</td>
<td>266</td>
</tr>
<tr>
<td>Ral 124</td>
<td>133</td>
</tr>
<tr>
<td>B.W.Africa 27</td>
<td>162</td>
</tr>
<tr>
<td>Mean</td>
<td>233</td>
</tr>
<tr>
<td><strong>Was</strong></td>
<td>172</td>
</tr>
<tr>
<td><strong>KA27</strong></td>
<td>261</td>
</tr>
<tr>
<td><strong>simulans:</strong></td>
<td></td>
</tr>
<tr>
<td>S.France 2</td>
<td>207</td>
</tr>
<tr>
<td>Japan Muna-1</td>
<td>176</td>
</tr>
<tr>
<td>Ral S-5</td>
<td>708</td>
</tr>
<tr>
<td>B.Congo S2</td>
<td>178</td>
</tr>
<tr>
<td>Mean</td>
<td>192</td>
</tr>
<tr>
<td>Ral 1-2</td>
<td>223</td>
</tr>
</tbody>
</table>
Alcohol Dehydrogenase Expression in Drosophila

**FIG. 5.—** Variation in ADH expression among pair-mated lines. Circles denote values for *melanogaster* lines, and black dots denote values for *simulans* lines. The three left-most plots depict the relationships between ADH activity and CRM level; the three right-most plots depict the relationships between CRM level and RNA level. The lines represent significant regressions estimated over all points, ignoring species except for CRM vs. RNA in larvae; in that case, the regression is for the *melanogaster* points only. For each regression line the slope is significantly different from zero, with *P* = 0.0001. R² values from the regression analyses are given on each plot.

can be considered representative. *Adh* clones have been obtained from these lines and are being used to investigate the molecular basis of the ADH expression differences between the species.

**Geographic Variation**

The eight isoallelic *simulans* lines from experiment 2, for which there was replication within geographic location, show a significant location effect on both adult activity (*P* < 0.05) and CRM (*P* < 0.001). The lines from France are high in activity (315 units), and those from South Africa are low (217), while Japan (241) and Tunisia (237) are intermediate. Significant location effects on adult activity (*P* < 0.05) and CRM (*P* < 0.01) were also observed in the pair-mated lines of *simulans* and *melanogaster* analyzed in experiment 3. In this case France is again high (359 units), and
West Africa is again low (232), while North America (307) and Japan (308) are intermediate.

Biochemical Basis of Activity Variation

Figures 4 and 5 show, for each of the stages analyzed in experiments 2 and 3, the plots of activity versus CRM level. In each case, there is, for each species analyzed separately, a significant regression of activity on CRM. An analysis of covariance was performed on the data set represented by each of these plots, to test for a between-species difference in the slope of the regression line. The slope difference was significant ($P = 0.046$) for only one of the five data sets: adults from the pair-mated lines. Since the slope difference in this case is very small and since the equivalent test for the isoallelic lines was not significant, we conclude that the relationship between activity and CRM is essentially the same for both species. Thus, the regression lines drawn in figures 4 and 5 are taken over all the data points, ignoring species. These plots show that most of the genetic variation in ADH activity, both within and between species, is accounted for by variation in the amount of ADH protein.

The observation that the activity difference between adults is due to a CRM-level difference is consistent with previous reports in the literature. Chambers et al. (1984b) and Heinstra et al. (1987) each analyzed one strain of each species and reported in adults a CRM-level difference that parallels the activity difference (in both cases the *melanogaster* strain was *Adh*). Dickinson et al. (1984) analyzed one *Adh* strain and one *simulans* strain and also found that in adults a CRM-level difference parallels the activity difference, which in this case was about fourfold.

Figures 4 and 5 also show the relationship, for experiments 2 and 3, between CRM level and RNA level. In only one case, *melanogaster* larvae in experiment 3, is there a significant regression of CRM on RNA. Thus, it appears that very little of the ADH-protein concentration variation among lines is accounted for by variation in ADH-mRNA.

Discussion

On the basis of a detailed developmental profile of one strain from each species—i.e., strains Was (*melanogaster*) and Ral 1-2 (*simulans*)—the patterns of change in ADH activity, in CRM level, and in RNA level are very similar for the two species. The only major difference in shape of the developmental profiles is due to the fact that, in adults between 0–1 and 7–10 d posteclosion, Was shows a large increase in both ADH activity and CRM whereas Ral 1-2 does not. In the relative abundance of distal and proximal transcripts through development, there is no apparent difference between these two strains.

Among strains within species there is, at all developmental stages, considerable quantitative variation in ADH activity, CRM level, and RNA level, but the only statistically significant species-specific differences are in pupal RNA level and in adult activity and CRM level. In both ADH activity and CRM level, in larvae and pupae, there are no significant differences between the species, and there is at these stages considerable overlap among the lines from each species. In adults *melanogaster Adh* strains have, on the average, about twice the ADH activity level of *simulans* strains, which is consistent with previous reports in the literature. As discussed in the Introduction, in laboratory tests *melanogaster* adults also consistently have a higher alcohol tolerance than do *simulans* adults. The ADH activity difference may cause or at least contribute to this alcohol-tolerance difference, although other genes could be at least
as important as those affecting ADH levels. A number of studies of variation within melanogaster have shown that ADH activity is not always correlated with alcohol tolerance (see Gibson and Oakshott 1982; Gibson and Wilks 1988). If ADH activity is a prime factor in determining alcohol tolerance, then the lack of a species difference in larval and pupal stages is difficult to reconcile with the suggestion that the two species show a difference in distribution with respect to alcohol-containing resources, since larvae and pupae are more intimately associated with the substrate than are adults. Nevertheless, it is possible that adult exposure to ethanol vapor during feeding and oviposition provides a selective agent that has influenced these species’ distributions.

Over strains within species, at each developmental stage there is a significant regression of ADH activity on CRM level, and this regression is homogeneous between species. Furthermore, the regressions have very high $R^2$ values. These results indicate that most, if not all, of the activity variation within and between species is accounted for by variation in the concentration of ADH protein. In contrast, the regression of CRM level on RNA level is not significant, except in melanogaster larvae. This result indicates that the ADH concentration variation among lines that is observed in pupae and adults is not accounted for by variation in mRNA template concentration. Therefore, the variation in ADH concentration must be due either to variation in rate of translation of the mRNA or to variation in protein stability. Both translational efficiency and protein stability might be due either to genetic background differences between the species (trans-acting modifier genes) or to cis-acting factors located within the Adh sequences that specify the mature message. Comparison of the melanogaster Adh$^5$ consensus sequence of Kreitman (1983) and the simulans Adh sequence of Cohn and Moore (1988) shows two amino acid differences, which could confer differential protein stability, and 19 other differences within the mature distal message, which could confer differential translational efficiency. We are presently conducting an interspecific gene transfer experiment, using P-element-mediated transformation, to determine the relative contributions of cis- and trans-acting factors to the Adh expression difference between adult of these two species.

Acknowledgments

We thank Jamie Bridgham and Justina Williams for their expert technical assistance. This work was supported by Public Health Service grant GM36388.

LITERATURE CITED


BRIAN CHARLESWORTH, reviewing editor

Received May 14, 1990; revision received July 11, 1990

Accepted August 3, 1990
Rates of DNA Change and Phylogeny from the DNA Sequences of the Alcohol Dehydrogenase Gene for Five Closely Related Species of Hawaiian Drosophila

Robert G. Rowan and John A. Hunt
Department of Genetics, University of Hawaii

The sequence of 1.6 kb of DNA surrounding the alcohol dehydrogenase (Adh) gene from five species of the Planitibia subgroup of the Hawaiian picture-winged Drosophila, with estimated divergence times of 0.4−5.1 Myr, has been determined. The gene trees which were found by using the sequence divergence from different regions of the sequences are generally in accord with the phylogeny proposed for these species when chromosomal inversions and island of origin are used. One of the species (D. picticornis) appears to be more distant from the other species in this group than they are from a member of the Grimshawi group (D. affinisdisjuncta) which is chromosomally more distant. Two of the species (D. differens and D. planitibia) show heterogeneity in the nucleotide changes in the Adh coding region, heterogeneity which is interpreted to be due to a gene conversion or recombination after hybridization between the two species. The minimal rate of nucleotide substitution of synonymous nucleotides and of nontranscribed nucleotides downstream from the coding region is estimated as $1.5 \times 10^{-8}$ and $1.1 \times 10^{-9}$ substitutions/nucleotide/year, respectively. This rate is two to three times the maximal rate estimated for mammalian synonymous substitutions.

Introduction

In previous work from our laboratory we examined the phylogenetic relationships of the planitibia subgroup of the picture-winged group of the Hawaiian Drosophila, by DNA hybridization and lowering of the melting temperature (Hunt et al. 1981; Hunt and Carson 1983) and by restriction mapping of the DNA surrounding the alcohol dehydrogenase gene (Bishop and Hunt 1988). The phylogenetic trees obtained from these two methods were in good agreement with each other and with the phylogenies proposed by others, who used chromosome inversions, enzyme polymorphism, morphology, and island of origin (Carson and Kaneshiro 1976; Carson and Yoon 1982).

Because of the sequential formation of the Hawaiian Island chain by volcanic action over a hot spot in the Pacific plate, it is possible to estimate the age of each island accurately by potassium/argon dating (McDougall 1979). The seven high islands in the Hawaiian chain form a chronological series in which each island is 0.5−1.5 Myr older than its adjacent eastern neighbor. Most species of the Hawaiian Drosophila are found only on one of the seven islands, and it is assumed that in most cases the species

1. Key words: Alcohol dehydrogenase, Hawaiian Drosophila, evolutionary rates, recombination.

2. Present address: Hopkins Marine Station, Stanford University, Department of Biological Sciences, Pacific Grove, California 93950

Address for correspondence and reprints: John A. Hunt, Department of Genetics, John A. Burns School of Medicine, University of Hawaii, 1560 East West Road, Honolulu, Hawaii 96822.

© 1991 by The University of Chicago. All rights reserved.
0737-4038/91/0801-0003$02.00
were formed, after the formation of the island, by colonization from an adjacent, older island. In many cases, when the phylogeny of a species group on an island is known, it is reasonable to assume that the oldest species diverged soon after the formation of that island, giving a more accurate dating for the divergence.

Both the technique of DNA hybridization and that of restriction mapping assume that the divergence between species results from the accumulation of single-base-pair substitutions. It is clear, however, from the examination of sequence comparisons of noncoding regions that many of the differences are caused by small deletions and insertions. Tajima and Nei (1984) have estimated that these differences are almost as great as the rate of synonymous changes in the first intron and 3’ coding region of the globin genes. When DNA hybridization is used to measure distances, there is a large difference between the ratio of lowering of the melting temperature and the amount of DNA which reassociates when comparing results between different phyla (Hunt et al. 1981; Caccone et al. 1988). This phenomenon could be caused either by the accumulation of small deletions and insertions or by unequal rates of accumulation of nucleotide changes as seen by Martin and Meyerowitz (1986). Werman et al. (1990) have shown that the DNA that fails to hybridize between Drosophila species does have some homology between the species and is not caused by large (>100–200-bp) deletions. However, because of the changing fraction of DNA that is measured by the hybridization technique, there is an inherent nonlinearity which would appear to be difficult to correct when one is comparing more distant species.

While the results from the different types of measurement for nuclear DNA in Hawaiian Drosophila have generally been in agreement (Johnson et al. 1975; Hunt and Carson 1983; Bishop and Hunt 1988), there is some discrepancy with the data from restriction-enzyme analysis of mitochondrial DNA (DeSalle and Giddings 1986). The mitochondrial phylogeny places the divergence of D. differens, D. planitibia, and D. silvestris/D. heteroneura as a sequential progression, while the nuclear DNA measurements indicate a bifurcation between the two pairs D. differens/D. planitibia and D. silvestris/D. heteroneura.

In the present study we have now extended the analysis of the five species of Hawaiian Drosophila—D. silvestris, D. heteroneura, D. differens, D. planitibia, and D. picticornis—by sequencing the DNA of the alcohol dehydrogenase (Adh) gene (Adh). The sequences analyzed include the complete coding region and almost 500 bp of DNA on the 3’ side of the poly(A) addition signal. This enables a comparison with the Adh region from several other Drosophila species, including another of the Hawaiian picture-winged group (D. affinidisjuncta; Rowan and Dickinson 1988) and D. melanogaster (Benyajati et al. 1981). The analysis of the sequence differences between the closely related species still cannot resolve the exact time of divergence of the two species D. differens and D. planitibia. Part of the reason is that between these species there is a nonuniformity in the distribution of nucleotide changes which is best explained by a gene conversion or recombination event between the two species. The heterogeneity covers most of the Adh coding region, which reduces the amount of the sequence which can be used for comparison of these two species. The suspected gene conversion or recombination is the first that has been described between two species and must have resulted from a hybridization between the species sometime in the recent evolutionary past.
Material and Methods

The isolation and restriction-site mapping of molecular recombinants of the Adh from Drosophila silvestris, D. heteroneura, D. dfferens, D. planitibia, and D. picticornis in bacteriophage lambda has been described elsewhere (Bishop and Hunt 1988). The restriction maps of the molecular recombinants used in the present work are shown in figure 1a. Blot hybridizations using Adh molecular recombinants from D. melanogaster (Goldberg 1980), hybridization with cDNA from larval poly(A) RNA from D. heteroneura (Bishop 1982; Bishop and Hunt 1988), and the DNA sequence of the Adh region from D. affinidisjuncta (Rowan and Dickinson 1988) locate the Adh protein coding sequence to a 1.6-kb EcoRI fragment in D. silvestris, D. heteroneura, D. dfferens, and D. planitibia. The homologous sequence is carried on a 4.5-kb EcoRI fragment in D. picticornis. For each species, the Adh-containing DNA fragments were cloned into the bacteriophage M13 mp18 and mp19 (Messing 1983; Yanisch-Perron et al. 1985) in preparation for dideoxynucleotide DNA sequencing (Sanger et al. 1977). The molecular recombinants were in turn subcloned after controlled deletion by Bal 31 exonuclease or by exonuclease III digestion (Henikoff 1984; Dale et al. 1985) and by further restriction-enzyme digestion, to provide a set of templates for DNA sequence determination. The distribution of the fragments is shown in figure 1b, together with the relationship of the coding and noncoding regions.

DNA sequences were read directly or by use of a semiautomatic gel reader using the Graf/Bar sonic digitizer (Science Accessories Corp.), from autoradiographs of denaturing polyacrylamide gels (Maxam and Gilbert 1977), by using an Apple Ile program (Komaromy and Govan 1984; J. A. Hunt, unpublished data) which produces files which can be used directly by the sequence analysis programs. Ambiguities were resolved by sequencing the uncertain region with avian myeloblastosis reverse transcriptase (Krawetz 1987), rather than with the Klenow fragment of Escherichia coli DNA polymerase and/or by sequencing the complementary DNA strand, when subclones containing the DNA sequence in this orientation were available. All sequences were confirmed by resequencing the templates at least once, and each autoradiograph was read at least three times. After comparison of Adh sequences from two species, apparent differences were verified by reading the relevant autoradiographs once again. Sequence data were collected and compared using the DNA and Protein Sequence Analysis (DPSA) program (Marck 1986) on an Apple Ile computer. More extensive similarity searches and alignment of the sequences of the different species—as well as comparisons with Adh sequences from other species—was done by using the SEQA and SEQH programs from Kanehisa (1984) on a VAX 11/750 VMS.

Results

Nucleotide Sequence

The nucleotide sequences of the five species of Hawaiian Drosophila determined in the present work, together with that of the same region from D. affinidisjuncta (Rowan and Dickinson 1988), are shown in figure 2. The single nucleotide changes that are shared by more than one species—and that are therefore phylogenetically informative—have been indicated by showing the changes in lowercase in figure 2. There are a total of 12 insertion or deletion events, ranging in length from 3 to 20 nucleotides in the six species, and there are five doublet and 10 single-base insertions or deletions. The alignments were first made by using the SFQA program from Ka-
FIG. 1.—a, Restriction maps of DNA inserts in bacteriophage molecular recombinant clones used for determination of sequences of Adh region. The region containing Adh is bounded by the EcoRI site at 5.4 kb and, except in the case of Drosophila picticornis, by another EcoRI site at 7.0 kb. S = SalI; H = HindIII; E = EcoRI; P = PstI; Het = D. heteroneura; Silv = D. silvestris; Plan = D. plantibida; Diff = D. differens; Pict = D. picticornis. b, Map of subcloned fragments containing Adh region of five species. The arrows indicate the regions of the DNA and the direction in which they were sequenced after subcloning into bacteriophage M13. The complete region was covered by using unidirectional deletions and further subcloning by restriction-enzyme digestion. The vertical bars on the arrows indicate regions that were subcloned by restriction-enzyme digestion and subcloned in opposite orientations to allow sequencing in both directions. The three exons which have been sequenced are shown by the filled bars above, and the transcribed region is shown by the heavy arrow.

nehisa (1984), and where the deletions and insertions in the sequences are obvious they were improved by eye. A possible duplication of a region of the second intron (residues 142–227 in D. heteroneura) in slightly different regions was revealed by another alignment program and is shown in figure 3. This indicates that, especially in the introns, it is not always possible to assume that a particular alignment algorithm has shown all of the best alignments—or, indeed, whether the putative changes shown occurred in evolution. This might lead to overestimation of the rate of nucleotide changes if such deletions are ignored. The sites used only for the construction of the parsimony trees are shown in summary in figure 4.

Distance Measurements

Because the distribution of nucleotide substitutions is not uniform across the region, the distances between the species were calculated for the following groupings in table 1: In the coding region, shown as the underscored region in figure 2, the
synonymous and nonsynonymous changes calculated by the method of Li et al. (1985) are shown in columns 1 and 3 of table 1. The fourfold-degenerate site changes calculated by the method of Lewontin (1989) are shown in column 2. In columns 4–7 the changes, calculated by the two-parameter method of Kimura (1980), are shown for the total coding region, for the 3' nontranscribed region following the poly(A) addition signals starting at position 1170 in D. heteroneura, for the total noncoding DNA, and for the total DNA of the 1.6-kb region.

Construction of Phylogenetic Trees

Gene trees were constructed from the distances by using several different algorithms. These are UPGMA with error estimates for the branching points (Nei et al. 1985), two programs from Felsenstein's PHYLIP package version 2.9 (Felsenstein 1988)—namely, (1) KITSCH, which assumes a clock, and (2) FITCH (Fitch and Margoliash 1967), which does not—and the neighbor-joining method of Saitou and Nei (1987). For these sequences the KITSCH and UPGMA programs give the same results. The same is true of the FITCH and neighbor-joining programs. In addition,
Fig. 2.—Comparison of sequences of Adh from five species of Hawaiian Drosophila [(HET), (SILV), (DIFF), (PLAN), and (PICT)] and from D. affinisdisjuncta from the grimshawi subgroup [(AFFI)]. Gaps were inserted to improve the alignment. The coding region is shown underlined at the beginning of each codon, and the putative poly(A) sites are shown as underlined italics. The dashes indicate nucleotides which are the same as in D. heteroneura, and changes are shown as the nucleotide. Phylogenetically informative nucleotide changes are shown in lowercase. The region of the potential crossover between D. di@rens and D. planitibiu covers the first 200 nucleotides. Abbreviations for the species are as in fig. 1.
FIG. 2. (Continued)
FIG. 2. (Continued)
FIG. 3.—First intron of Adh (bp 143–227), showing alignment different from that shown in fig. 2. Several duplications can be inferred, the first from the repetition of the TTATTATTA sequence (region 1) from Drosophila affinisjuncta in the two species pairs D. heteroneura/D. silvestris and D. differens/D. planitibia; this sequence is not repeated in D. picticornis. The second inferable duplication is of the TTTTTT (region 2) from the D. heteroneura/D. silvestris and D. differens/D. planitibia pairs, which is repeated in D. affinisjuncta and as TATTTT in D. picticornis. The AATAA repetition (region 3) is found in all of the species with some changes, including deletions of single bases, except in the D. differens/D. planitibia pair, where it is deleted. On the basis of the scheme shown in this figure, there are two fewer nucleotide changes for the heteroneura versus affinisjuncta and planitibia and one fewer for heteroneura versus picticornis. The other comparisons have the same number of changes. The alignments shown in this figure assume a major role of slippage in replication.

FIG. 4.—Alignment and position of all phylogenetically informative apparent single-nucleotide changes between species. The positions are those from Drosophila heteroneura. Where gaps occur, the position indicated is that of the first heteroneura nucleotide that follows the deletion. Abbreviations for the species are as in fig. 1.
<table>
<thead>
<tr>
<th>SPECIES PAIR</th>
<th>CODING REGION</th>
<th>NONCODING REGION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Synonymousa</td>
<td>Nonsynonymousa</td>
</tr>
<tr>
<td>Heteromeura/silvestris</td>
<td>0.00 (0.00)</td>
<td>0.003 ± 0.002</td>
</tr>
<tr>
<td>Heteromeura/differens</td>
<td>0.023 ± 0.012 (0.01)</td>
<td>0.014 ± 0.005</td>
</tr>
<tr>
<td>Heteromeura/planitibia</td>
<td>0.006 ± 0.006 (0.00)</td>
<td>0.012 ± 0.005</td>
</tr>
<tr>
<td>Heteromeura/picticornis</td>
<td>0.153 ± 0.032 (0.08)</td>
<td>0.020 ± 0.006</td>
</tr>
<tr>
<td>Heteromeura/affinisjuncta</td>
<td>0.127 ± 0.029 (0.09)</td>
<td>0.012 ± 0.005</td>
</tr>
<tr>
<td>Silvestris/differens</td>
<td>0.023 ± 0.012 (0.01)</td>
<td>0.014 ± 0.005</td>
</tr>
<tr>
<td>Silvestris/planitibia</td>
<td>0.006 ± 0.006 (0.00)</td>
<td>0.012 ± 0.005</td>
</tr>
<tr>
<td>Silvestris/picticornis</td>
<td>0.153 ± 0.032 (0.09)</td>
<td>0.023 ± 0.006</td>
</tr>
<tr>
<td>Silvestris/affinisjuncta</td>
<td>0.128 ± 0.029 (0.09)</td>
<td>0.016 ± 0.005</td>
</tr>
<tr>
<td>Differentis/planitibia</td>
<td>0.017 ± 0.010 (0.01)</td>
<td>0.002 ± 0.002</td>
</tr>
<tr>
<td>Differentis/picticornis</td>
<td>0.166 ± 0.034 (0.10)</td>
<td>0.023 ± 0.006</td>
</tr>
<tr>
<td>Differentis/affinisjuncta</td>
<td>0.141 ± 0.030 (0.01)</td>
<td>0.039 ± 0.004</td>
</tr>
<tr>
<td>Planitibia/picticornis</td>
<td>0.145 ± 0.031 (0.10)</td>
<td>0.022 ± 0.006</td>
</tr>
<tr>
<td>Planitibia/affinisjuncta</td>
<td>0.121 ± 0.028 (0.09)</td>
<td>0.037 ± 0.003</td>
</tr>
<tr>
<td>Picticornis/affinisjuncta</td>
<td>0.172 ± 0.034 (0.10)</td>
<td>0.022 ± 0.006</td>
</tr>
<tr>
<td>Heteromeura/melanogaster</td>
<td>1.15 ± 0.16 (0.95)</td>
<td>0.140 ± 0.017</td>
</tr>
<tr>
<td>Silvestris/melanogaster</td>
<td>1.16 ± 0.16 (0.95)</td>
<td>0.140 ± 0.017</td>
</tr>
<tr>
<td>Differentis/melanogaster</td>
<td>1.19 ± 0.16 (1.04)</td>
<td>0.135 ± 0.017</td>
</tr>
<tr>
<td>Planitibia/melanogaster</td>
<td>1.13 ± 0.15 (0.99)</td>
<td>0.135 ± 0.017</td>
</tr>
<tr>
<td>Picticornis/melanogaster</td>
<td>1.20 ± 0.17 (1.02)</td>
<td>0.129 ± 0.016</td>
</tr>
<tr>
<td>Affinisjuncta/melanogaster</td>
<td>1.12 ± 0.16 (1.16)</td>
<td>0.138 ± 0.017</td>
</tr>
</tbody>
</table>

a Distances calculated according to the method of Li et al. (1985), without using any correction for codon bias. The numbers in parentheses are distances for fourfold degenerate synonymous sites and are calculated as hits per synonymous codon (Lewontin 1989); in all cases, 42% of the synonymous sites were in this class, and 85–107 codons are represented.

b Distances calculated according to the method of Kimura (1980). ND = not determined (because regions could not be aligned for comparisons with D. melanogaster).
all of the phylogenetically informative state changes, nucleotide changes, and single-base deletions and insertions were used for construction of trees by the DNA-parsimony and compatibility methods from PHYLIP. No attempt was made to use the deletions and insertions larger than one base, since only two phylogenetically informative changes were found that were shared by more than two species; and there is no way that such deletions and insertions can be properly weighted or fit to the parsimony program, since they are nonrevertible and do not fit the criteria for use in this program.

A summary of the trees obtained for the data is shown in figures 5 and 6. The neighbor-joining trees shown in figure 6 indicate that there is some inequality of rates in each arm as compared with the equivalent UPGMA trees in figure 5. However, when the errors in determining the branch points as shown in figure 5 are taken into account, these differences are not significant. The total noncoding trees in figures 5A and 6A and the 3' nontranscribed tree in figure 5B indicate that there is a bifurcation between the *D. heteroneura/D. silvestris* and *D. differens/D. planitibia* pairs. The synonymous codon trees in figures 5C and 6C indicate that the divergence of these species is in the order *D. differens, D. planitibia, and D. heteroneura/D. silvestris*. The nonsynonymous (amino acid) trees shown in figures 5D and 6D are very similar to those shown in the total noncoding region, except that *D. affinidisjuncta*, the supposed outgroup species, is now clustered with *D. differens* and *D. planitibia*. This is not found in the other trees (A, B, and C; figs. 5 and 6), where the divergence of *D. affinidisjuncta* is very close to that of *D. picticornis*. The trees found by using the parsimony method are topologically the same as those obtained by using the neighbor-joining method (fig. 6).

![Diagram showing relationships of six species of Hawaiian *Drosophila*](image_url)

**Fig. 5.**—UPGMA dendrograms showing relationships of six species of Hawaiian *Drosophila*, on the basis of distances expressed as substitutions/nucleotide/Myr, and calculated from total noncoding (A), 3' nontranscribed (B), synonymous (C), and nonsynonymous (D) regions of *Adh* region. These dendrograms were calculated by using UPGMA with error calculation (Nei et al. 1985). The distances are from table 1. The horizontal lines are proportional to the distances, and the vertical lines have no distance component. The error bars represent 1 SE on either side of the branch point. Abbreviations for the species are as in fig. 1.
Evidence for Recombination within Adh of Drosophila differens and D. planitibia

There is a distinct difference in the distances calculated between D. differens and D. planitibia, depending on whether the 3' nontranscribed region or the 5' region including the whole of the coding region is used for the comparison. Evidence that this heterogeneity is due to hybridization between the two species—with subsequent recombination, or gene conversion—is seen by the excess of the shared nucleotide changes in the first 200 nucleotides at the 5' end of the sequence as compared with the 3' nontranscribed region. The distribution of the unique, shared nucleotide changes in the three species pairs D. heteroneura/D. silvestris, D. differens/D. planitibia, and D. affinisdisjuncta/D. picticornis is shown in figure 7, as is the distribution of the changes found between the species pairs. No such heterogeneity of distribution of changes is found for the other species pairs. In addition, the deletion at position 182–194, which is unique to D. differens and D. planitibia, is in the region where a recombination event is suspected. Unfortunately, there is too little information to apply either Stephens’s (1985) or Sawyer’s (1989) statistical test for gene conversion.

Determination of Rate of Nucleotide Substitution between Species

To estimate the rate of nucleotide change we can use the branch points in the phylogeny and the approximate time when these branches occurred. The dating of the emergence of the Hawaiian Islands has been estimated, with a standard deviation of <10% (McDouggall 1979). Previous considerations of the phylogeny and biogeographic distribution of the planitibia subgroup make it possible to estimate the place and maximal time of their divergence.

Drosophila heteroneura and D. silvestris diverged after the formation of the island...
Fig. 7.—Distribution of nucleotide changes along sequence of Adh region. The fraction of the total changes is shown for each of the three pairs of species in 200-bp regions along the DNA. A, Distribution of changes between pairs. B, Distribution of nucleotides shared only between pairs of species. ■ = silvestris/heteroneura; □ = differens/planitibia; ▧ = picticornis/affinisjuncta.

of Hawaii 0.4 Myr ago (Mya). If the divergence of D. planitibia and D. differens is ignored, because of the heterogeneity of the rates of change over the sequences being compared, the second branch point is between the D. heteroneura/D. silvestris pair and D. differens. The maximal time of divergence for these species would be from the time of formation of Molokai (1.8 Mya). However, the data from mitochondrial DNA of other species, from Maui, suggest that the divergence of those planitibia subgroup species which share the Xr inversion was not the primary event and could
have occurred at the time of the formation of Maui (1.3 Mya) (DeSalle and Templeton 1988).

From the chromosomal phylogeny it is expected that the divergence of *D. picticornis* would be the next branch point and that this branch would have occurred at the time of formation of the island of Oahu (3.5 Mya) and that *D. affinisjuncta*, which is a member of the *grimshawi* subgroup, possibly would have diverged at the time of formation of Kauai (5.1 Mya). Because all of the trees indicate that *D. affinisjuncta* is closer to the other *planitibia* group species than to *D. picticornis*, this latter assumption may not be true. There are representatives of the *grimshawi* group on Kauai. Therefore, it is more likely that the time of divergence of these species was some time after the formation of Kauai and before the formation of Oahu.

The plots of the nucleotide distances from the coding and noncoding regions are shown in figure 8, where either the Kauai or the Oahu formation time is used as the divergence time for *D. picticornis* and where the Maui and Hawaii formation times are used as the other two divergence times. The line drawn through the points is a regression line using these three nodes. The estimated rates of nucleotide change, as measured by the regression lines, are shown in table 2. On the basis of the rates derived from a divergence for *D. picticornis* on Kauai, the times of divergence of *D. affinisjuncta* and *D. planitibia* are 3.9 ± 0.44 Mya and 0.67 ± 0.09 Mya, respectively. No measure of the within-species nucleotide polymorphism has been made by sequencing the DNA from these regions on other chromosomes, so the rates cannot be corrected for this polymorphism.

**Discussion**

Rate of DNA divergence in Hawaiian *Drosophila* Species

The slopes of the regression lines in figure 8 are twice the rate of nucleotide change. These rates are shown in table 2. In most cases the correlation is greater when the divergence time of *D. picticornis* is taken as 5.1 Mya (the age of the oldest lava flows on Kauai; McDougall 1979) than when the divergence is taken as 3.5 Mya (the time of emergence of Oahu). The maximum rate when the slope of the regression lines from the Kauai divergence is used is $1.7 \times 10^{-8}$ substitutions/nucleotide/year; however, there is on the regression line a negative intercept which may exaggerate the true slope. The next highest rate, $1.0 \times 10^{-8}$ substitutions/nucleotide/year, is that

<table>
<thead>
<tr>
<th>DNA REGION COMPARED</th>
<th>DIVERGENCE ON KAUI</th>
<th>DIVERGENCE ON OAHU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean ± SE Slope</td>
<td>Coefficient of</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correlation Rate*</td>
</tr>
<tr>
<td>Synonymous ..........</td>
<td>0.034 ± 0.002</td>
<td>0.996</td>
</tr>
<tr>
<td>Synonymous, zero origin</td>
<td>0.030</td>
<td>...</td>
</tr>
<tr>
<td>3' End (noncoding)</td>
<td>0.020 ± 0.0003</td>
<td>0.999</td>
</tr>
<tr>
<td>Total coding</td>
<td>0.010 ± 0.0009</td>
<td>0.985</td>
</tr>
<tr>
<td>Total noncoding</td>
<td>0.019 ± 0.0006</td>
<td>0.999</td>
</tr>
<tr>
<td>Total nucleotides</td>
<td>0.014 ± 0.0004</td>
<td>0.994</td>
</tr>
</tbody>
</table>

* Calculated, as substitutions/nucleotide/Myr, from the regression of the nucleotide distances with island age, from the data in fig. 8.
Fig. 8.—Regression lines of distances (as substitutions/nucleotide) between species, plotted against estimated time of divergence of main groups of species. The divergence times shown are 0.4 Myr for heteroneura and silvestris (●); 1.3 Myr for differens and the heteroneura/silvestris pair (●); and either 3.5 Myr (a and c) or 5.1 Myr (b and d) for the divergence of picticornis vs. the rest of the species (○). In a and b the circles and squares represent the total noncoding regions and the total coding regions, respectively; In c and d the circles and squares represent the synonymous and 3' nontranscribed regions, respectively.

for the 3' nontranscribed region which we expect to be free of any selection constraints, since it comes after the poly(A) addition signal.

While no measurement of polymorphism within species has yet been made by DNA sequence analysis, measurements made using restriction-enzyme-site polymorphism indicated that the degree of polymorphism was similar in each of the species. The mean level of polymorphism for the region covering 10 kb surrounding the Adh gene was 0.6%/nucleotide (Bishop and Hunt 1988). Estimates from recombinant DNA clones from different regions on the 3' side of the Adh range from 3%/nucleotide for D. picticornis to 1.5%/nucleotide for D. silvestris (Bishop and Hunt 1988; K. A. Houtchens and J. A. Hunt, unpublished data). If the level of polymorphism in the
Adh region is similar in all of the species, then the slope of the regression line would not be affected by the polymorphism.

Because the regression line for the synonymous changes does not pass through or above the origin (fig. 8d), the slope may be anomalously high. A better estimate for the synonymous rate can be found by taking the average distances between D. picticornis and the other species and assuming a time of divergence of 5.1 Mya. In this case a rate of $1.5 \times 10^{-8}$ substitutions/nucleotide/year is found (table 2). This is comparable to the rate found for the 3' nontranscribed region ($1.0 \times 10^{-8}$ substitutions/nucleotide/year). Previous studies of distances in the Hawaiian Drosophila have found rates of $0.5 \times 10^{-8}$ substitutions/nucleotide/year for restriction-enzyme analysis and DNA hybridization (Hunt and Carson 1983; Bishop and Hunt 1988). The rate found for the total DNA of the Adh region, $0.75 \times 10^{-8}$ substitutions/nucleotide/year, is quite close to this value but may reflect that only a small region of the genome is being surveyed by this method. Data from other gene regions of the same species may help to determine whether there is indeed a variation of rates over the genome. It is possible that the synonymous rate may change by a factor of two (Shields et al. 1988), and it is quite possible that a significant portion of the noncoding DNA is subject to some form of selection and that therefore its rate will be lower.

Comparison with Rates of Change in Other Species

The major difficulty in determining rates of nucleotide change in most insect species is the lack of any good fossil record for the determination of dates of species divergence. Where attempts have been made to determine rates, the most frequent reference for divergence time is Throckmorton (1975). Here the age of a species is inferred from geographical distribution of species and from the estimated time of continental drift (for a more comprehensive discussion of the available information, see Beverley and Wilson 1984). The major inference is that the divergence of the virilis and melanogaster groups occurred >40 Mya and probably <60 Mya, provided that there was no migration across the Atlantic or Pacific Oceans. Rates of synonymous substitutions based on this divergence time for Drosophila are given as $0.5-1.6 \times 10^{-8}$ substitutions/nucleotide/year (Moriyama 1987). However, it should be noted that the Hawaiian Drosophila are assumed to have migrated from continental land masses, from which the Hawaiian Islands are >4,000 km apart. This is not much less than the shortest distance between the South American continent and Africa. If we consider that the nucleotide distance found between the Hawaiian Drosophila and D. melanogaster is a measure of the divergence time of the Drosophila and Sophophora divergence, then, when the synonymous rate of change of $1.5 \times 10^{-8}$ substitutions/nucleotide/year is used, the estimated time of divergence is 41 Mya. However, this may be complicated by the apparently slow rate of change of synonymous nucleotides in D. melanogaster (Shields et al. 1988), a slow rate which leads to an underestimation of the time of divergence. A very similar estimate is obtained when the nonsynonymous rates are used. Further information on rate heterogeneity awaits the use of an outgroup to the Drosophila and Sophophora groups.

Li et al. (1985) have measured the rate of synonymous nucleotide divergence of mammalian and avian species. The rates found in their study are $0.14-1.18 \times 10^{-8}$ substitutions/nucleotide/year, with a mean of $0.47 \times 10^{-8}$ substitutions/nucleotide/year. In rodents the maximal rate of change is found to be $0.8 \times 10^{-8}$ substitutions/nucleotide/year for synonymous substitutions (Sharp and Li 1989). It appears that the rate of synonymous nucleotide substitution in Drosophila is two to three times
higher than the average found in mammals. The *Drosophila*-to-mammalian rate ratio is a minimum value because the fossil evidence gives a minimum time since divergence, whereas the island-dating method gives a maximum time since divergence. A similar conclusion was made by Moriyama (1987) by using the synonymous substitution rate from *Drosophila Adh* and heat-shock proteins.

Deletion and Insertional Changes

There are many changes between these species that cannot be accounted for by single nucleotide substitutions. These changes vary from the large insertions and deletions which are caused by transposing elements and are readily detected by restriction mapping (Aquadro et al. 1988) to small deletions and insertions. These latter range in size, from three nucleotides to >200 nucleotides, with the majority being <100 nucleotides. This latter class would be largely undetected by restriction mapping, unless four base-cutting enzymes are used (Kreitman and Aguadé 1986). The insertions and deletions are generally limited to the noncoding region. While deletions and insertions are seen in the coding regions of the multigene chorion family (Jones and Kafatos 1982), very few changes of this type seem to be tolerated in the single-copy gene coding sequences.

In the *Adh* coding region, there is a deletion of six nucleotides (two amino acids) between the *Sophophora* and *Drosophila*—but no other. Of the >10-nucleotides deletions that are found in the Hawaiian *Drosophila*, only two are phylogenetically informative in that they are shared by two or more species. One of these two, the deletion shared by *D. differens* and *D. planitibia* in the first intron, could be accounted for by gene conversion or recombination between the species. The other, just after the 3' end of the coding region, is shared between *D. silvestris* and *D. heteroneura* and appears to be either a deletion and insertion or two different insertions. Of the five deletions or insertions found singly in the different species, two are found in each of the species *D. affinisdisjuncta* and *D. picticornis*. Most of the changes cannot be accounted for by a slippage-type error in DNA duplication, except for the insertion in *D. affinisdisjuncta* in the second intron and, possibly, the duplications seen in the first intron (fig. 3). In the other deletions there are no common features—such as repetitive sequences or short repeats flanking the insertions—which would indicate that they have occurred by the process of transposition and excision, and in this respect these deletions do not follow the types of change, found by Jones and Kafatos (1982) in the chorion gene sequences of the silk moth, which are caused mostly by simple sequence changes.

The comparison of the 3' noncoding regions of *D. heteroneura* and *D. picticornis* shows that 9% of the sequence over this region differs because of deletions and insertions. Nucleotide substitutions account for 12% of the variation in the rest of the region. A similar result is found for the ratio of deletions and insertions in the two regions having differing rates of change and that are close to the glue genes of *D. melanogaster*, *D. orena*, and *D. yakuba* (Martin and Meyerowitz 1986).

Comparison between *Adh* Regions of Other Species

*Adh* is one of the most frequently sequenced DNA regions of the *Drosophila* genus. The phylogenetic relationships of this region have been intensively studied in the *melanogaster* subgroup (Kreitman 1983; Bodmer and Ashburner 1984; Cohn et al. 1984; Eastal and Oakeshott 1985; Stephens and Nei 1985; Coyne and Kreitman 1986). In the case of such closely related species as *D. sechellia*, *D. simulans*, and *D. mauritiana*, it is extremely difficult to infer an unambiguous phylogeny from the
sequence of the Adh region, since the amount of polymorphism within the species is very close to the between-species differences. The distances involved are similar to the distances which we find here for the species D. silvestris, D. heteroneura, D. differens, and D. planitibia. In both the melanogaster and planitibia species groups, anomalous mitochondrial DNA phenotypes are found that would indicate that introgression has occurred between the species (DeSalle and Giddings 1986; Solignac et al. 1986).

The codon usage in the Adh of the Hawaiian species is different from that of D. melanogaster Adh. This was shown by doing a likelihood-ratio test comparing Adh codon usage in D. melanogaster and D. heteroneura, by using the program Monte Carlo 2×N written by W. Engels (personal communication) and based on the procedure described by Lewontin and Felsenstein (1965). The Monte Carlo simulation using identical marginal totals as the observed table showed a probability of identity of the species of <0.004. The same test within the Hawaiian species had a probability of identity of ~0.99.

The distribution of codon usage is heterogeneous, as is found in D. melanogaster (Ashburner et al. 1984). The difference in the “scaled” $\chi^2$ values (Shields et al. 1988) found for the two species groups also indicates a difference in the codon usage. Starmer and Sullivan (1989) have also found that the usage of the third base in coding regions is different between the Sophophora and Drosophila species groups including D. affinisdisjuncta.

If the rates of nucleotide change are equal, the ratio of transversions to transitions is expected to be 2:1, and hence the fraction of transitions is expected to be 0.33. The fraction of transitions for all of the planitibia group species (minus D. picticornis) is 0.66 ($\chi^2 = 66.4$, 1 df, $P < 0.01$), and for D. picticornis versus the rest of the species the ratio is 0.56 ($\chi^2 = 45.3$, 1 df, $P < 0.01$). In contrast, if one uses the noncoding regions, the fraction of transitions for all of the closely related species (minus D. picticornis) is 0.52 ($\chi^2 = 60.4$, 1 df, $P < 0.01$), and for D. picticornis versus the rest of the species the ratio is 0.46 ($\chi^2 = 30.9$, 1 df, $P < 0.01$). There is a definite reduction in the fraction of transitions as the time of divergence between the species increases. This is to be expected, since transversions cannot be reversed as readily as transitions, a fact which allows an accumulation of transversions over longer periods of evolutionary time.

Phylogenetic Consequences of Between-Species Recombination

The construction of species trees by using nucleotide sequence differences is confounded by the presence of polymorphism in the ancestral populations, which makes the gene trees that are constructed less informative for closely related species than might be expected (Pamilo and Nei 1988). A similar problem certainly exists with phenotypic characters used for determining species trees, but polymorphism can be measured more accurately for nucleotide changes. A further limitation lies in the possibility of hybridization between D. planitibia and D. differens, with a consequent recombination or gene conversion within the Adh region. In this case the region of the DNA involved in the recombination must be excluded from the measurements used to determine the gene trees for the species. This reduces the number of nucleotides that can be compared between the two species in determining their distances. Interspecific hybridization should also affect other nuclear genes and the measurements made by DNA hybridization. For measurements using mitochondrial DNA, the effects could be even more drastic, as is seen by the apparent replacement, in certain popu-
lations, of one mitochondrial DNA species by another (Powell 1983; DeSalle et al. 1986; Solignac et al. 1986).

There is virtually no difference between the gene trees that are derived from the data by using UPGMA (fig. 5) and those that are derived from the data by using the neighbor-joining method (fig. 6), all of the differences being in the relative placement of *D. diffrers* and *D. planitibia*. For these species the greatest discrepancies lie in the coding regions where the gene conversion is predicted, as is shown by the comparisons of synonymous and nonsynonymous changes. These anomalies are also seen in the parsimony trees.

If we consider all of the previous trees determined by distance methods (Johnson et al. 1975; Hunt and Carson 1983; DeSalle and Giddings 1986), it would seem that *D. planitibia* is always closer to *D. diffrers* than it is to either *D. heteroneura* or *D. silvestris*, even for the mitochondrial DNA. But the differences in most cases are not significant. In the trees shown here, this phylogeny is confirmed (figs. 5a and 5b). The only exception derives from the coding regions where *D. planitibia* appears to be closer to *D. heteroneura* and *D. silvestris*. More information from different genes may help to determine whether this is indeed the case.

Duplication of Adh Region

In the *mulleri* group there is both a duplication of the active Adh and an additional pseudogene adjacent to the duplication (Fischer and Maniatis 1985; Atkinson et al. 1988). In both the *melanogaster* subgroup and *D. pseudoobscura* there is an apparent duplication of Adh, but the duplicated gene appears to have a different function (Schaeffer and Aquadro 1987). In *D. affinnisjuncta* Rowan and Dickinson (1988) have shown that there is a duplication of almost 500 bp on the 5' side of the first exon, which duplication is found approximately 1.6 kb downstream from this point. The 1.6-kb subclones which we have sequenced appear to terminate in this duplication. In *D. heteroneura*, where a fragment of 300 bp has been sequenced downstream from the 1.6-kb *EcoRI* fragment, evidence for the duplication is found (L. Brezinsky and J. A. Hunt, unpublished data). This duplication may be the remnant of a gene duplication, but, since it stops just at the beginning of the exon, it clearly has no present function. Preliminary evidence from other species outside the picture-winged group of Hawaiian *Drosophila* indicates that this duplication is not present in species more distant than 10 Mya (R. H. Thomas and J. A. Hunt, unpublished data).

Sequence Availability

These sequences have been deposited in GenBank under accession numbers M36781 (*heteroneura*), M36782 (*picticornis*), M36783 (*planitibia*), M36784 (*silvestris*), and M36785 (*diffrers*).

Acknowledgments

We thank Drs. G. Carmody, I. Felger, J. Martinez-Cruzado, and R. Thomas for critically reading the manuscript; Dr. T. Lyttle for statistical advice; and Dr. H. Carson for his continued interest in the project and for his unlimited knowledge of the evolution of the Hawaiian *Drosophila*. This work was supported by grants BSR 82-14338 and BSR 85-21810 from the National Science Foundation.
LITERATURE CITED


MASATOSHI NEI, reviewing editor

Received January 9, 1990; revision received August 10, 1990

Accepted August 16, 1990
Ribosomal DNA Variation Within and Between Species of Rodents, with Emphasis on the Genus *Onychomys*

Marc W. Allard* † and Rodney L. Honeycutt†

*Department of Organismic and Evolutionary Biology, Harvard University, and
†Department of Wildlife and Fisheries Sciences, Texas A&M University

Patterns of restriction-endonuclease site and length variation at the nuclear rDNA locus (18S + 28S rRNA gene complex) were examined in rodents. Of the 164 restriction sites mapped for seven species, 22 were conserved (mapping to the 18S, 28S, and 5.8S genes and ITS1) in all three *Onychomys* species as well as in *Mus musculus* and in three closely related peromyscine rodents, *Peromyscus boylii, P. eremicus,* and *Reithrodontomys megalotis.* The nontranscribed spacer (NTS) region revealed most of the variation among these taxa, with the patterns of variation grouping into the following categories, (1) intraindividual variation revealing as many as four site-specific repeat types within an individual, (2) intraspecific and interspecific site variation confined to the NTS, and (3) length variation in both the transcribed and NTS regions. Length variation in the 28S rRNA gene was also examined in 17 additional rodent species, and most size differences mapped to the divergent domain, D8, found in sequence comparisons between *Mus* and *Rattus.*

The systematic implications of rDNA variation are discussed using the perspective gained from these rodent comparisons.

Introduction

The rDNA repeating unit in vertebrates consists of three gene regions (18S, 5.8S, and 28S rRNA genes), two internal transcribed spacers (ITS-1 and ITS-2) which separate each gene, and an external transcribed spacer (ETS) at the 5' end of the 18S rRNA gene (fig. 1). This gene complex is represented in the mammalian genome by multiple copies (50–250) which are located on one to several pairs of chromosomes (Appels and Honeycutt 1986). Each copy or repeating unit is flanked by a nontranscribed spacer (NTS) which has promoter activity and considerable size variation in some vertebrates and invertebrates (Treco et al. 1982).

In recent years the patterns and processes of rDNA evolution have come under close scrutiny, and several observations have been made concerning the evolution of this gene system. First, sequence heterogeneity (as assessed by restriction-site variation) in coding and noncoding regions of rDNA repeats is very low within some species, presumably as the result of concerted evolution (Dover 1982; Arnheim 1983). The process of concerted evolution reduces sequence variation via non-Mendelian mechanisms (unequal exchange, gene conversion, and sequence transposition), resulting in relatively homogeneous sequences within a species. Second, most species examined for rDNA variation reveal both within-individual and between-individual variation, and the distribution and frequency of rDNA variants demonstrate a definite geographic pattern (Suzuki et al. 1986, 1987). This level of within-individual and between-in-
dividual variation presumably reflects a lag time in turnover of variation, resulting from the influx of new variation by mutation and from the fixation of that variation by non-Mendelian processes. Third, patterns of rDNA variation have been used to evaluate the phylogenetic relationships among taxa of several vertebrate groups (Hillis and Davis 1986; Larson and Wilson 1989; Mindell and Honeycutt 1989).

Although a considerable amount of information exists on rDNA structure and evolution, few detailed comparative studies of mammalian rDNA variation have been conducted, with the possible exception of phylogenetic studies on hominoid primates (Wilson et al. 1984; Schmickel et al. 1990). Our objective in the present study is to investigate rDNA variation among rodent species. We have approached this problem by evaluating variation at several taxonomic levels of divergence, including intraindividual, intraspecific, intrageneric, and intergeneric diversity, and in most cases the comparisons have involved restriction-endonuclease-site variation in peromyscine and murine rodents. Existing nucleotide sequence data for *Mus* and *Rattus* (from GenBank; Bilofsky and Burks 1988), together with restriction-site data, have been used to evaluate the types of change, magnitude of change, and overall patterns of divergence seen at the rodent rDNA locus.

**Material and Methods**

**Specimens Examined**

Detailed patterns of intraindividual, intraspecific, and interspecific variation were examined in *Onychomys* species collected from the following localities (sample size in parenthesis): *O. leucogaster*—Oregon, Morrow Co. (two); New Mexico, Bernalillo Co. (one), DeBaca Co. (one), Luna Co. (two), and Sandoval Co. (one); and Texas, Lamb Co. (five) and Winkler Co. (three); *O. torridus*—Arizona, Pinal Co. (two); and *O. arenicola*—New Mexico, Eddy Co. (one), Luna Co. (one), Socorro Co. (one), and Valencia Co. (two). In addition to these *Onychomys* genera, intergeneric comparisons included one individual of each of the following species: *Mus musculus*—Harvard Biological laboratory stock; *Rattus norvegicus*—Harvard Biological laboratory stock; *Peromyscus alstoni*—Mexico, Tlaxcala; *P. bandeanus*—Mexico, Jalisco; *P. boylii*—Arizona, Pima Co.; *P. eremicus*—Arizona, Pima Co.; *P. maniculatus*—Washington; *P. nudipes*—Costa Rica; *P. thomasi*—Mexico, Oaxaca; *P. pirrensii*—Panama, Darien prov.; *Reithrodontomys megalotis*—New Mexico, Luna Co.; *Reithrodontomys montanus*—Texas, Lubbock Co.; *Ochrotomys nutalli*—Arkansas, Pulaski Co.; *Neotomys sumichrasti*—Mexico, Jalisco; *Nectomys parvipes*—Suriname, Marowijne; *Sigmodon hispidus*—Texas, Lubbock Co.; *Neotoma micropus*—New Mexico, Luna Co.; and *Ototylomys phyllotis*—Costa Rica.

**FIG. 1.**—Schematic of rRNA gene complex for rodents, including representative regions of N1S, E1S, ITS, and three gene coding regions—18S, 5.8S, and 28S rRNA. p2546 and p119 define the regions covered by the probes used in restriction-site mapping. Conserved restriction-endonuclease sites found in rodent taxa are shown at the top. Restriction-site maps include all geographic variation observed within each species. Restriction sites above/below the maps were conservative/variable sites within that species. All sites were shared between the two *O. torridus* samples examined. Triangles represent presumed insertion/deletion events. Mapped sites affected by the hypothesized insertion/deletion events were aligned with respect to one another. Mapped location of 60 sites used in the analysis of *Onychomys, Peromyscus,* and *Reithrodontomys* are shown at the bottom, and these correspond to the 60 sites listed in table 1. Restriction endonucleases mapped are denoted by single letters as follows: A = ApaII; B = BamHI; L = BclI; G = BglII; T = BstEII; C = Clal; D = DraI; E = EcoRI; H = HincII; N = HindIII; K = KpnI; P = PstI; V = PvuII; S = SacI; F = SalI; X = XbaI; O = XhoI.
FIG. 2.—Insertion/deletion events observed in 28S gene around SacI (S) and PvuII (V) restriction sites. Symbols are as in fig. 1. Numbers indicate the size (in bp) of insertion/deletions. Taxa sharing the *Onychomys* pattern include *Peromyscus maniculatus*, *P. boylii*, *P. eremicus*, *P. banderanus*, *P. nudipes*, *P. alstoni*, *P. pirrensis*, *Reithrodontomys montanus*, *R. megalotus*, and *Neotoma*.

Restriction-Endonuclease Mapping

High-molecular-weight DNA was isolated using a modified procedure of Bingham et al. (1981). Approximately 3–5 μg of nuclear DNA was digested with restriction endonucleases according to the manufacturer's specifications (New England Biolabs). The digested DNAs were electrophoresed in a 0.8% agarose gel in 1 × TAE (0.04 M Tris, 0.005 M sodium acetate, 0.001 M disodium ethylenediamine tetraacetate, pH
Ribosomal Gene Variation in Rodents 75

8.2) buffer. The gel was stained with ethidium bromide, and the DNA was transferred to a nylon membrane (GeneScreen Plus; DuPont) by using the Southern (1975) blot procedure, with some modifications for alkaline transfer (Chomczynski and Qasba 1984).

Two rDNA probes were used to map the 18S and 28S rRNA gene complex. These probes were p19, a 4.8-kb insert containing part of the Mus 28S rRNA gene, and p2546, a 1.9-kb fragment containing most of the Mus 18S rRNA gene (fig. 1; Arnheim 1979). All probes were radiolabeled using the random priming method (Feinberg and Vogelstein 1984). After hybridization with a specific probe, the nylon filters were washed (three times at 42°C with 2 × saline sodium citrate, 0.1% sodium dodecyl sulfate; twice at 60°C in 0.1% saline sodium citrate, 0.1% sodium dodecyl sulfate) and were exposed to X-ray film for 18–48 h. Prior to hybridization with a second probe, the old probe was stripped from the filters, and the filters were prehybridized. Both rRNA probes were hybridized separately to all filters examined.

Seventeen restriction endonucleases (for list, see fig. 1) were used to map the 18S and 28S rDNA complex for Onychomys arenicola, O. leucogaster, O. torridus, P. boylii, P. eremicus, Reithrodontomys megalotus, and M. musculus. In addition, Rattus norvegicus, P. alstoni, P. banderanus, P. maniculatus, P. thomasi, P. pirenensis, Reithrodontomys montanus, Ochrotomys nuttalli, Nyctomys sumichrasti, Neotomys parvipes, S. hispidus, Neotoma micropus, and Ototylomys phyllotis were mapped for length variation in the 28S rRNA gene only. Mapping was accomplished by single and double digests (Nathans and Smith 1975). Of the 136 possible combinations of double digests for the 17 restriction endonucleases used, we performed 52 double digests for each of three species of Onychomys. In addition, Mus controls were run on each gel as an internal comparison vis-à-vis the known sequence of the Mus rDNA repeat. Therefore, equal sized fragments could be confidently oriented to known mapped and sequenced data. Double digests were chosen so as to minimize the distance between a known site within the region of the probe and an unknown site that we wished to map accurately. Geographic variation was assessed mainly by comparison with the known mapped standard for a particular species.

Phylogenetic relationships within the genus Onychomys and among the genera Onychomys, Peromyscus, and Reithrodontomys were determined by parsimony analysis outlined in Swofford's PAUP (version 3.0, option branch and bound) and Farris's Hennig 86 (version 1.5, option ie*) computer programs. Each restriction site was treated as an unordered character and was scored as present or absent. The use of heterologous probes confined to the transcribed regions, plus the large size of the mammalian NTS, increases the probability that some restriction sites in the NTS will be undetectable. These sites were coded as “missing data” in the character-state matrix and were used in the parsimony analysis. Mus musculus was used as an outgroup in an effort to determine the polarity of characters in the phylogenetic tree. The criterion for comparing trees derived from parsimony analysis is based on tree length, with the most parsimonious tree being the shortest tree. When two or more trees of equal length were found, a strict consensus tree was constructed for all these trees combined.

Results
Patterns of Variation
Restriction Sites

From the 164 restriction sites mapped in the present study, the rRNA gene coding regions were shown to be highly conservative in all taxa (fig. 1). For instance, of the 21 sites mapped in the gene regions, no variation was observed in the 5.8S or 28S...
Table 1
Coding for Phylogenetically Informative Restriction-Endonuclease Sites for Onychomys, Peromyscus, Reithrodontomys, and Mus

<table>
<thead>
<tr>
<th>SPECIES</th>
<th>BglII</th>
<th>XbaI</th>
<th>KpnI</th>
<th>BglII</th>
<th>HindIII</th>
<th>PvuII</th>
<th>EcoRI</th>
<th>Clal</th>
<th>Sall</th>
<th>KpnI</th>
<th>BellI</th>
<th>SacI</th>
<th>ApaLI</th>
<th>PstI</th>
</tr>
</thead>
<tbody>
<tr>
<td>O. leucogaster:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texas</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>New Mexico</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>Oregon</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>O. arenicola, New Mexico</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>O. torridus, Arizona</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>P. boylii</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>P. eremicus</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R. megalolis</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>M. musculus</td>
<td>-</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>

*NOTE.*—Data are expressed in binary style: 0 = absence; 1 = presence; - = unknown.

*Numbers 1-60 in column heads are the same as those in Fig. 1.

Genes, and only one variable site (Saci) was found in the 5' end of the 18s rRNA gene. The SacI site was absent in Peromyscus boylii.

The ITS regions revealed more variation, with nine sites differing among genera. By far the most variable region was the NTS, with more than half of the 132 variable sites being unique to individual species. Most of the NTS variation (intraspecific, interspecific, and intergeneric) involved the regions extending 4.5 kb from the 5' end of the 18s gene and 3.0 kb from the 3' end of the 28s gene (fig. 1, sites 10-26 and 30-48, respectively).

**Length Variation**

Overall repeat length varied from 45.5 kb to 48 kb, with at least three regions of the rDNA repeat contributing to this variation. One of the major sources of variation was the NTS (fig. 1). In this region, species of Onychomys and Mus musculus revealed size increases of 2 kb and 650 bp, respectively. Onychomys leucogaster and O. arenicola also revealed a 50-bp addition in the ITS-1. Finally, the size of the 28s rRNA gene varied in length among most rodent genera. A detailed examination of 21 rodent
ribosomal gene variation in rodents

species revealed size variation mapping to the 28s gene 3' end, between Sac1 and PvuII sites (fig. 2).

**Intraindividual Variation**

Nine restriction endonucleases revealed multiple repeat types within individuals of *Onychomys*. In all cases except one (*Sac1* site 29 in ITS-2), this variation mapped to the NTS. The greatest number of repeat types within an individual was four, and the average number of repeat types for *O. leucogaster*, *O. torridus*, and *O. arenicola* was 1.64, 1.4, and 1.6, respectively. Some repeat types were unique to a species, but several were shared between species. Four polymorphic sites were shared between *O. arenicola* and *O. leucogaster* (fig. 1, sites 29, 36, 39, and 59).

**Phylogenetic Analysis**

A phylogenetic analysis using rDNA site variation was conducted for the three peromyscine rodent genera, *Onychomys, Peromyscus, and Reithrodontomys* (table 1). Analysis of the 60 phylogenetically informative characters, including the polymorphic sites (all sites were located in the NTS, except for the three sites 27–29 in the ITS), produced 1,500 equally parsimonious trees of length 101 and with a consistency index (CI) of 0.59. The strict consensus tree (fig. 3) constructed from these data revealed the following: (1) Texas and New Mexico *O. leucogaster* form a clade with respect to Oregon *O. leucogaster* (2) *Onychomys torridus* is more closely related to *O. leucogaster*.
than is *O. arenicola*, with trees depicting either an *O. arenicola*/*O. leucogaster* or an
*O. arenicola*/*O. torridus* relationship being five and six steps longer, respectively. The
*Onychomys* relationships are maintained when a parsimony analysis is conducted
without an outgroup, as well as when either *Peromyscus* or *Reithrodontomys* is used
as an outgroup rather than *Mus*. (3) The genus *Peromyscus* is paraphyletic, with
*P. boylii* and *P. eremicus* grouping closer to *Reithrodontomys megalotis*.

Patterns of nucleotide sequence change for four regions of the rDNA complex
were examined at several levels of evolutionary divergence by using two approaches.
First, equation (8) of Nei and Li (1979) was used to calculate $\delta$ (nucleotide divergence
per nucleotide site) for all mapped sites in *Onychomys*, *Peromyscus*, and
*Reithrodontomys*. As can be seen in figure 4, $\delta$ increases as taxonomic divergence increases, with
the NTS and ITS regions reflecting a faster evolutionary rate as compared with that
in the coding region. Second, a divergence profile was constructed using known nu-
cleotide sequence data for *Mus* and *Rattus*. A similar trend relative to the rate of
divergence depicted by $\delta$ values can be seen for regions of the rRNA gene complex.

Discussion

Length Variation

The length variation observed in the rDNA repeat is comparable to that seen in
other organisms (Appels and Honeycutt 1986). One major cause of rDNA length
variation in mammals is the deletion/addition of repeats mapped to regions in the NTS, either 3' to the 28S rRNA gene or 5' to the 18S rRNA gene (Erickson and Schmickel 1985; Appels and Honeycutt 1986; Suzuki et al. 1986). In rodents and humans this length variation has been found at several levels of divergence, including intraindividual, intraspecific, and interspecific. Although the exact location of repeat-length variation has not been determined for most species of rodents examined in the present study, interspecific repeat-length differences are apparent and map to the general region of the NTS. The accurate mapping of length variation can be seen in Onychomys (fig. 1), in which length differences occur 14 kb 5' to the 18S rRNA gene. This region is somewhat removed from those indicated for MUS (Kuehn and Amheim 1983) and Homo (Erickson and Schmickel 1985). As can be seen in figure 1, the major difference in length between MUS and the other rodents (a 650-bp addition in MUS) does map to the vicinity of the variable region reported by Kuehn and Arnheim (1983). The ITS also demonstrates length variation between species (e.g., Mus vs. Rattus; Michot et al. 1983), and at least some ITS length variation occurs between species of Onychomys (fig. 1), with the remaining rodent taxa being constant in ITS size.

Insertions and deletions within the 28S gene of several vertebrates and other eukaryotes have been reported (Hillis and Davis 1987; Hancock and Dover 1988; Mindell and Honeycutt 1989). These regions of length variation have been termed “expansion segments” or “divergent domains” and represent the major cause of size
difference in the large rRNA gene of eukaryotes (Clark et al. 1984; Hassouna et al. 1984). On the basis of nucleotide sequence comparisons, 12 divergent domains have been located in eukaryotes, with two domains (D2 and D8) revealing most of the size variation between the rodent genera Mus and Rattus (Hassouna et al. 1984). The Mus 28S gene (Hassouna et al. 1984) has two SacI sites bracketing four divergent domains, D7–D10, with a PvuII site mapping within D8. These restriction-endonuclease sites were used to localize deletion/insertion differences among additional rodent 28S genes (fig. 2), and in most cases the size variation maps close to the D8 region found in Mus and Rattus.

The location and size of specific additions/deletions in the 28S gene may provide some phylogenetic information for deriving relationships among rodent genera. Carleton’s (1980) systematic treatment of neotomine-peromyscine rodent genera suggests a closer relationship among the genera Peromyscus, Reithrodontomys, Onychomys, Ochrotomys, and Neotoma, with Ototylomys being more distantly related. As can be seen in figure 2, Ototylomys shares deletions/insertions with Mus and Rattus. Thus, the pattern seen in most peromyscine-neotomine rodents can be considered synapomorphic, with Ochrotomys and P. thomasi possessing uniquely derived types.

Although there seems to be some phylogenetic information relative to 28S deletions/insertions, one must interpret restriction-site maps with considerable caution. While no difference in the size of the 28S rRNA gene could be distinguished between Mus and Rattus when Southern blotting was used, at the nucleotide-sequence level these same regions revealed a 17–31-bp difference between the genera (Hassouna et al. 1984). Therefore, in the absence of direct sequencing, there are limits to one’s ability to assume homology (in an operational sense) between insertions/deletions of similar size.

Site Variation

Patterns of restriction-site variation at the rDNA locus have been used to determine both relationships among species and the patterns of geographic and population variation within species (Suzuki et al. 1987; Mindell and Honeycutt 1989). In most comparative studies the NTS and ITS regions have been shown to be the most variable
Ribosomal Gene Variation in Rodents

Fig. 4.—Nei and Li's (1979) nucleotide substitutions per nucleotide site, for subregions of rRNA gene complex compared at intraspecific, interspecific, and intergeneric levels by using 86 restriction-endonuclease sites, with no missing data, from peromyscine rodents and Mus. The number of pairwise comparisons were as follows: 116 intraspecific, 115 intrageneric, and 69 intergeneric. Intergeneric distance values are also compared between Mus and Rattus at the nucleotide level from published sequence data (Bilofsky and Burks 1988). For restriction-site values, NTS and ETS were lumped, because of the uncertain location of the site for transcription initiation in the grasshopper mouse, and the ITS regions were grouped to increase sample size.

(Appels and Honeycutt 1986), and, as can be seen both in the restriction-site maps (fig. 1) and in the divergence estimates (fig. 4), these two regions are the most divergent in rodents as well. The utility of these regions for resolving phylogenetic relationships among rodents, is equivocal, however. Multiple repeat types within and between in-
individuals are quite common, and most variation of this type maps to the NTS (La Volpe et al. 1985; Seperack et al. 1988). Similar variation was observed in the present study, and in some cases repeat types defined by a particular restriction site were even shared between species. Within a species, the frequency of particular repeat types does demonstrate patterns of geographic variation similar to those derived from other characters (Suzuki et al. 1986, 1987), and the NTS variation seen in *Onychomys* is no exception. For instance, the relationships among different geographic samples of *O. leucogaster* (fig. 3) are congruent with relationships derived from morphological and mitochondrial DNA data (Riddle and Honeycutt 1990), suggesting the utility of the NTS for deriving regional relationships among populations within a species.

Restriction-site variation in the NTS and ITS is less clear with regard to the derivation of phylogenetic relationships among species of peromyscine rodents. First, the close relationship, depicted by the rDNA data, between *O. torridus* and *O. leucogaster* relative to *O. arenicola* is incongruent with the results of most other comparative studies. Allozymes, morphology, and mitochondrial DNA suggest a sister-group relationship between *O. leucogaster* and *O. arenicola* (Sullivan et al. 1986; Riddle and Honeycutt 1990). If one were to accept the rDNA phylogeny based on NTS variation, then the patterns of variation seen with these other characters would be more complicated. Chromosomal data, however, generally support the rDNA results, but the chromosomal differences separating the three species of *Onychomys* present some difficulties with respect to determining character-state polarity (Baker and Barnett 1981). For instance, all chromosomal differences among the three species are the result of additions/deletions of constitutive heterochromatin to/from the short arms of autosomes, and it is impossible to evaluate whether these additions/deletions represent independent events or events resulting from common ancestry. Second, the intergeneric comparisons are also equivocal. The monophyly of *Onychomys* and the closer relationship between *Reithrodontomys* and *Peromyscus* are congruent with the inferences to be drawn from other comparative data sets (Hooper and Musser 1964; Carleton 1980). Nevertheless, the paraphyly demonstrated for the genus *Peromyscus* is not congruent with other data (Koop et al. 1984). Even in Carleton’s review (Carleton 1980), where *Peromyscus* is considered paraphyletic, *P. boylii* and *P. eremicus* were placed in the same clade.

The comparisons of rodent taxa examined in the present study span a time period of ~4–15 Mya or more (Carleton and Eshelman 1979; Jacobs and Pilbeam 1980; Flynn et al. 1985). The overall results of these comparisons suggest that restriction-site variation at the rDNA locus may be of limited utility in systematic studies of rodents (especially in studies involving intergeneric comparisons), with the evolutionary rates of coding regions being too slow and those of the NTS being too fast. Therefore, we suggest that further studies on rodent rDNA variation be conducted at the nucleotide sequence level and be confined to the more variable regions mapped to the 28S gene or, possibly, to the ITS region.

Acknowledgments

Special thanks to Robert J. Baker, The Museum, Texas Tech University, for frozen tissue from many of the rodent taxa and to Kim Nelson for specimen collection and laboratory assistance. We are also grateful to Ron Adkins, Mike Arnold, Jim Carpenter, Scott Davis, Brian Hanks, David Hillis, Michael Lynch, Masatoshi Nei, David Pilbeam, MaryEllen Ruvolo, Ron Van Den Bussche, Ward Wheeler, Scott Williams, and two anonymous reviewers for comments on earlier drafts of the manuscript. This research was supported by National Science Foundation grant BSR
85084479 to R.L.H. and by grants-in-aid, to M.W.A., from the American Society of Mammalogists and Sigma Xi, The Scientific Research Society. During the course of this research M.W.A. was supported by NIH Genetics Training grant GM07620.

LITERATURE CITED


MASATOSHI NEI, reviewing editor

Received March 29, 1990; revision received July 26, 1990

Accepted July 26, 1990
Variation in Mitochondrial Cytochrome b Sequence in Natural Populations of South American Akodontine Rodents (Muridae: Sigmodontinae)

Margaret F. Smith and James L. Patton
Museum of Vertebrate Zoology, University of California, Berkeley

A 401-bp fragment of the mitochondrial cytochrome b gene was sequenced from polymerase chain reaction–amplified products for 20 natural populations representing 12 species of South American akodontine rodents (Muridae). Variation among these taxa increased with their hierarchical position, from comparisons within local populations to those among different genera. Two individuals from the same local population differed by < 1% sequence divergence. Sequence divergence among geographic samples within a species was 0.25%–8%, while that among species was 3%–21%. Comparisons of the akodontine sequences with that for the house mouse show 21%–25% sequence difference. A parsimony-based phylogenetic analysis of the data supports the placement of the taxon Microxus within Akodon (sensu stricto), of Bolomys just outside the Akodon cluster, and of Chroeomyys as a separate genus quite distinct from the other members of this group. This phylogenetic hypothesis is identical to that determined from electrophoretic data but is quite divergent from the present taxonomy of the group.

Introduction

The genus Akodon is one of the largest, most complex, and taxonomically most poorly understood taxa of South American murid rodents. The current combination of Akodon and related genera to form the tribe Akodontini is summarized in table 1. Systematic analyses of this complex group are limited to morphological, cytogenetic, and/or electrophoretic treatments of only selected subsets of species (e.g., see Apfelbaum and Reig 1989; Myers 1989; Myers and Patton 1989; Patton et al. 1989; Myers et al. 1990).

Here we report on variation in a 401-bp sequence of the mitochondrial cytochrome b gene (cyt b) in 20 natural populations representing 12 species of Akodon and related taxa. Levels of variation are compared within local populations, between geographic samples within species, and among different species and genera. The sequence data are used in a phylogenetic analysis to illustrate the degree of resolution provided by these data at various hierarchical levels within this taxonomic complex. For general review of the uses of mtDNA in evolutionary studies, see the work of Wilson et al. (1985), Avise (1986), Avise et al. (1987), Moritz et al. (1987), and Harrison (1989).

Patton et al. (1989) analyzed electrophoretic data on six genera within the tribe Akodontini, including most taxa examined here. Electrophoretic data from additional populations and species supplement this set to provide a series of samples that matches

1. Key words: polymerase chain reaction, mitochondrial DNA, cytochrome b, akodontine rodents.

Address for correspondence and reprints: Margaret F. Smith, Museum of Vertebrate Zoology, 1120 Life Sciences Building, University of California, Berkeley, California 94720.
Table 1
Supraspecific Groups of South American Akodontine Rodents
(tribe Akodontini, subfamily Sigmodontinae)

<table>
<thead>
<tr>
<th>Genus</th>
<th>Approximate No. of Species</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Akodon</strong></td>
<td></td>
</tr>
<tr>
<td><em>Akodon</em></td>
<td>25</td>
</tr>
<tr>
<td><em>Abrothrix</em></td>
<td>6</td>
</tr>
<tr>
<td><em>Deltamys</em></td>
<td>1</td>
</tr>
<tr>
<td><em>Hypsimys</em></td>
<td>2</td>
</tr>
<tr>
<td><em>Chroeomys</em></td>
<td>1</td>
</tr>
<tr>
<td><em>Bolomys</em></td>
<td>6</td>
</tr>
<tr>
<td><em>Microxus</em></td>
<td>3</td>
</tr>
<tr>
<td><em>Oxymycterus</em></td>
<td>9</td>
</tr>
<tr>
<td><em>Chelemys</em></td>
<td>4</td>
</tr>
<tr>
<td><em>Geoxus</em></td>
<td>1</td>
</tr>
<tr>
<td><em>Notiomys</em></td>
<td>1</td>
</tr>
<tr>
<td><em>Blarinomys</em></td>
<td>1</td>
</tr>
<tr>
<td><em>Podoxymys</em></td>
<td>1</td>
</tr>
<tr>
<td><em>Lenoxus</em></td>
<td>1</td>
</tr>
<tr>
<td><em>Juscelinomys</em></td>
<td>1</td>
</tr>
</tbody>
</table>

**NOTE.**—Table follows Reig (1987), with modifications by Pearson (1984) and Myers and Patton (1989).

Those used for the mitochondrial DNA (mtDNA) analysis. The electrophoretic loci reflect changes in the nuclear genome, whereas the cyt b sequences represent the maternally inherited mitochondrial genome. Thus, the comparison between the results from these two data sets allows one to examine the potential confounding effects of lineage sorting in the clonally inherited mtDNA.

**Material and Methods**

**Electrophoretic Data**

Twenty-one enzymes and other proteins encoded by 26 presumptive structural gene loci were examined using horizontal starch-gel electrophoresis. Aqueous extracts of kidney were used for all systems. The loci examined and the gel-running conditions have been reported by Patton et al. (1989). Data, taken from Patton et al. (1989), for 12 populations from 10 different species are used here. To supplement these data, a total of 215 additional individuals from nine new populations of the previously studied species and from two additional species are also incorporated in this analysis. Allele frequencies for these populations are available from the authors on request. The BIOSYS-1 computer package (Swofford and Selander 1981) was used to construct a phylogenetic tree by using the Wagner distance algorithm (Farris 1972) with Rogers’s (1972) measure of genetic distance.

**mtDNA Sequence Data**

DNA was extracted from frozen liver tissue by using the sodium dodecyl sulfate-proteinase K/phenol/RNAsen method (Maniatis et al. 1982). Sequence was obtained for a segment of cyt b, one of 13 protein-coding genes in the circular mitochondrial genome (for review, see Brown 1985). Primers are designated by a number in a series
Fig. 1.—Autoradiogram illustrating sequence from mtDNA cyt $b$ gene in akodontine rodents. A 482-bp fragment of the mtDNA cyt $b$ gene was amplified with primer pair MVZ 05-MVZ 04. Sequence is shown here from seven species, run on a 6% polyacrylamide gel. Numbered sequences are as follows: geographic variation within *Akodon subfuscus* MVZ 174109 (lane 1), MVZ 172970 (lane 2), and MVZ 174229 (lane 3); geographic variation within *A. torques* MVZ 174053 (lane 4) and MVZ 171720 (lane 5); geographic variation within *A. aerosus* MVZ 172870 (lane 6), MVZ 171679 (lane 7), and MVZ 172818 (lane 8); *A. boliviensis* MVZ 171607 (lane 9); variation within a local population of *Bolomys amoenus* MVZ 172878 (lane 10) and MVZ 172879 (lane 11); *A. andinus* MVZ 174063 (lane 12); *A. (Chrooemys) jelskii* MVZ 173085 skin extract (lane 13) from the same population as for MVZ 173083 and MVZ 173084. The numbers along the sides indicate the position of that base in the complete mtDNA sequence of the house mouse (Bibb et al. 1981). The sequences shown are of the light strand, with the 5' end at the bottom.

from the laboratory of the Museum of Vertebrate Zoology (MVZ) (University of California, Berkeley). Primer MVZ 04 employed in the present study was a slightly edited form of a “universal” primer designed by Kocher et al. (1989) for a segment of the cyt $b$ gene. Primer MVZ 05 was based on a primer designed by S. Pääbo. The site recognized by MVZ 05 is in the tRNA gene for glutamic acid, adjacent to the cyt $b$ gene. In the following sequences the letters in parentheses identify the light (i.e., L) or heavy (i.e., H) strand of mtDNA, and the numbers following the letters give the position of the 3' base of the primer in the complete mtDNA sequence for the house mouse (Bibb et al. 1981): MVZ 05(L14115) 5'-CGAAGCTTGATATGAAAAAC-
Fig. 2 (Continued)
MUS musculus
A. boliviensis
171607

A. per
171612
A. subfuscus
172569+172970

A. juninensis
173038+173039
A. mollis
173057+173058
A. torques
173214

A. aerosus
174053+174054

A. andinus
174062+174063

Microxus mimus
171748+171744

A. andinus
173073+173074

Fig. 2 (Continued)
Fig. 2 (Continued)
FIG. 2.—Nucleotide sequence from mtDNA cyt b gene. Sequence is reported for 401 bases from 20 populations representing 12 species of South American akodontine rodents. The dots indicate bases that are the same as in the Mus reference sequence on the first line. In 13 of the populations both individuals examined had identical sequences; these are reported on a single line. In six populations the second individual differed at one or more sites, and the variable sites only are shown on a second line. The sequences are of the light strand. The corresponding sequence for the house mouse (Bibb et al. 1981), beginning at site 14139 and ending at site 14540, is given for reference.
Table 2
Silent and Replacement Changes for Pairwise Comparisons at
401 Sites Within and Among the 12 Species of Akodontines

<table>
<thead>
<tr>
<th>NO. (%) in Codon Position</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transitions:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silent</td>
<td>629 (7.0)</td>
<td>102 (1.1)</td>
<td>2,345 (25.9)</td>
</tr>
<tr>
<td>Replacement</td>
<td>2,060 (22.8)</td>
<td>1,598 (17.7)</td>
<td>0</td>
</tr>
<tr>
<td>Transversions:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silent</td>
<td>19 (0.2)</td>
<td>99 (1.1)</td>
<td>449 (5.0)</td>
</tr>
<tr>
<td>Replacement</td>
<td>872 (9.6)</td>
<td>859 (9.5)</td>
<td>18 (0.2)</td>
</tr>
</tbody>
</table>

CATCGTTG-3'; MVZ 04(H14542) 5'-GCAGCCCCTCAGAATGATATTTGT-CCTC-3'. MVZ 05 in combination with MVZ 04 amplifies a 426-bp segment of the cyt b gene. With the primers added at each end, the total length of the amplified fragment is 482 bp.

Amplification of double-stranded product was performed in 25 μl or 50 μl total reaction volume with 27–35 cycles of the polymerase chain reaction (PCR) with the *Thermus aquaticus* DNA polymerase (Saiki et al. 1986, 1988). Amplification was performed either with a Perkin Elmer–Cetus thermal cycler with denaturation at 93°C for 30 s, annealing at 60°C for 30 s, and extension at 72°C for 1 min or with a Techne Programmable Dri-Block PHC-1 with denaturation at 93°C or 94°C for 1.5 min, annealing at 35°C–60°C for 1.5 min, and extension at 72°C for 2 min.

Electrophoresis of 5 μl of the double-stranded amplification product was run in a 4% NuSieve agarose minigel in Tris-borate-ethylenediaminetetraacetate buffer (Maniatis et al. 1982). The gel was stained with ethidium bromide, and the product bands were visualized by fluorescence under ultraviolet light. Agarose plugs (~4 μl) containing the double-stranded product were diluted 1:100 in glass-distilled water and were melted at 65°C for 10 min. The resulting template was then used in a PCR reaction with unbalanced priming to generate single-stranded product (Gyllensten and Erlich 1988), using an annealing temperature of 35°C–60°C. Primers were used in a ratio of 1:10, 1:20, 1:50, or 1:100, with 1:50 being the most generally successful ratio. The reaction mixture containing single-stranded product was subjected to three cycles of centrifugation dialysis (Centricon 30; Amicon Corp.), and 7 μl was used as a template for sequencing with Sequenase 2 (U.S. Biochemical Co. protocol) by following the Sanger method (Sanger et al. 1977). The sequencing primer used was the limiting primer in the unbalanced amplification reaction. Both strands were amplified for all individuals, in order to score sequence close to both primers. Scoring from the two strands overlapped in the middle by 9–114 bases. DNAs from most individuals were amplified and sequenced more than once. A further check on scoring was provided by comparison of the sequences of the two individuals from each local population.

Sequences were entered into BIONET (IntelliGenetics, Inc. 1988) for alignment and translation. Sequence data were analyzed using PHYLIP (version 3.2; Felsenstein 1989) run on an IBM-PC/AT. Cladistic analyses of character-state matrices by using parsimony methods were performed using DNAPARS, with a majority-rule consensus tree produced by CONSENSE; a majority-rule consensus tree based on 500 bootstrap replicates was produced using DNABOOT. The parsimony method used by Felsenstein
FIG. 3.—Hierarchical levels of sequence divergence. a, Hierarchical levels of mtDNA cyt b sequence difference within and among natural populations of akodontine rodents, plus comparisons with house mouse (from Bibb et al. 1981), based on 401 bp. b, Hierarchical levels of mtDNA cyt b sequence difference, similar to 3a, but with each transversion counted 10 times as much as each transition.

in DNAPARS and DNABOOT is based on the work of Eck and Dayhoff (1966) and on the work of Kluge and Farris (1969) and uses the method of Fitch (1971) to count the number of base changes needed on a given tree. The method assumes that sites evolve independently, as do lineages, that the ancestral base at a site is not known, that base substitutions are improbable over the time span involved, and that the amounts of evolution in the various lineages are fairly equal (see Felsenstein 1989, p. 201). The consensus tree from a bootstrap analysis “can be considered to be an overall
Table 3
Sequence Differences (transitions/transversions) for mtDNA Cytochrome b Gene

<table>
<thead>
<tr>
<th>Population</th>
<th>MVZ Number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Mus*</td>
<td>56/37</td>
</tr>
<tr>
<td>2. Akodon boliviensis</td>
<td>57/39</td>
</tr>
<tr>
<td>3. A. puer</td>
<td>63/40</td>
</tr>
<tr>
<td>4. A. subfuscus</td>
<td>62/40</td>
</tr>
<tr>
<td>5.</td>
<td>62/40</td>
</tr>
<tr>
<td>6.</td>
<td>54/39</td>
</tr>
<tr>
<td>7. A. kofordi</td>
<td>56/39</td>
</tr>
<tr>
<td>8. A. juninensis</td>
<td>43/43</td>
</tr>
<tr>
<td>9. A. mollis</td>
<td>11. A. torques</td>
</tr>
<tr>
<td>12. A. aerosus</td>
<td>174110</td>
</tr>
<tr>
<td>13. A. subfuscus</td>
<td>174229</td>
</tr>
<tr>
<td>14. A. aerius</td>
<td>174429</td>
</tr>
<tr>
<td>15. Microxus minus</td>
<td>172870</td>
</tr>
<tr>
<td>16. A. juninensis</td>
<td>172878</td>
</tr>
<tr>
<td>17. A. torques</td>
<td>172849</td>
</tr>
<tr>
<td>18. Bolomys amoenus</td>
<td>172818</td>
</tr>
<tr>
<td>19. A. andinus</td>
<td>172818</td>
</tr>
<tr>
<td>20. A. (Chroeomys) jelskii</td>
<td>172849</td>
</tr>
<tr>
<td>21. Microxus minus</td>
<td>172870</td>
</tr>
</tbody>
</table>

Note.—Pairwise comparisons of the number of sequence differences in 401 bp of the mtDNA cyt b gene for 12 species of akodontines and for the house mouse. The diagonal shows within-population variation, i.e., the specimen listed compared with a second individual from the same population. For population 21 the second individual of A. (Chroeomys) jelskii had three ambiguities, so the comparison is based on 398 bp.

*Data are from Bibb et al. (1981).

The estimate of the phylogeny” (Felsenstein 1985, p. 786). The initial parsimony analyses were unweighted. An additional bootstrap analysis was run using only phylogenetically informative transversions, by recoding the data set with only the two character states: purine and pyrimidine. All data matrices were prepared by following suggestions of Swofford (1985, pp. 3-11-3-12) for DNA data sets, including eliminating invariant sites, excluding identical taxa, and (optionally) including only phylogenetically informative sites where at least two bases each occur in more than one taxon. The amino acid sequence data were analyzed with the protein-parsimony (PROTPARS) procedure in PHYLIP, which uses a parsimony method “intermediate between Eck and Dayhoff’s (1966) method of allowing transitions between all amino acids and counting those, and Fitch’s (1971) method of counting the number of nucleotide changes that would be needed to evolve the protein sequence” (Felsenstein 1989, p. 163). A majority-rule consensus tree for the amino acid data was produced with the CONSENSE program in PHYLIP. *Mus musculus* was designated as the outgroup for all trees.

Specimens Examined

For each locality the sample size for the electrophoretic analysis is given in parentheses, followed by the museum catalog numbers of the voucher specimens for the individuals that were sequenced. Vouchers are cataloged in the mammal collection of the MVZ.
<table>
<thead>
<tr>
<th>POPULATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 11 12 13 14 15 16 17 18 19 20 21</td>
</tr>
<tr>
<td>47/44 47/43 52/43 48/46 58/45 57/45 48/41 50/41 51/41 51/38 48/44 47/45</td>
</tr>
<tr>
<td>37/7 35/6 39/6 33/9 33/10 32/10 34/4 34/4 41/10 42/29 46/33 46/34</td>
</tr>
<tr>
<td>33/5 31/4 37/4 34/7 32/8 31/8 30/4 32/4 38/10 37/29 45/33 46/34</td>
</tr>
<tr>
<td>39/6 39/5 43/5 44/8 38/9 37/9 40/5 42/5 40/11 37/28 50/32 47/33</td>
</tr>
<tr>
<td>38/6 38/5 42/5 43/8 37/9 36/9 39/5 41/5 39/11 38/28 49/32 46/33</td>
</tr>
<tr>
<td>39/6 39/5 41/5 42/8 36/9 35/9 36/5 38/5 38/11 36/28 48/32 45/33</td>
</tr>
<tr>
<td>37/7 33/6 39/6 37/7 35/8 34/8 32/4 34/4 35/8 45/29 43/33 43/34</td>
</tr>
<tr>
<td>43/5 38/4 43/4 38/7 32/8 33/8 37/2 37/2 42/8 40/27 51/31 48/32</td>
</tr>
<tr>
<td>14/1 19/2 30/0 22/5 23/4 22/4 35/4 37/4 38/12 47/27 40/31 38/32</td>
</tr>
<tr>
<td>1/0 11/1 28/1 22/6 20/5 19/5 31/5 33/5 44/13 50/28 45/32 45/33</td>
</tr>
<tr>
<td>0/0 26/2 23/5 22/6 21/6 34/4 36/4 46/12 48/29 46/31 48/32</td>
</tr>
<tr>
<td>0/0 17/5 25/4 24/4 36/4 38/4 45/12 44/27 42/31 44/32</td>
</tr>
<tr>
<td>3/0 16/3 17/3 31/7 32/7 46/11 43/30 44/30 43/31</td>
</tr>
<tr>
<td>0/0 1/0 30/8 31/8 42/10 49/27 47/29 44/30</td>
</tr>
<tr>
<td>1/0 29/8 30/8 41/10 48/27 46/29 43/30</td>
</tr>
<tr>
<td>0/0 2/0 41/10 46/29 47/29 46/30</td>
</tr>
<tr>
<td>0/0 43/10 46/29 48/29 47/30</td>
</tr>
<tr>
<td>1/0 50/27 51/31 47/30</td>
</tr>
<tr>
<td>0/0 42/10 43/11</td>
</tr>
<tr>
<td>0/0 19/1</td>
</tr>
<tr>
<td>3/0</td>
</tr>
</tbody>
</table>

Akodon (Akodon)

*aerosus.*—Peru: Depto. Junin, 10 km WSW San Ramon, 1,275 m (N = 6; MVZ 172870 and MVZ 172871); Depto. Cusco, 72 km NE Paucartambo, 1,460 m (N = 28; MVZ 171679 and MVZ 171680); Depto. Puno, 4 km NNE Ollachea, 2,380 m (N = 44; MVZ 172818 and MVZ 172819); 14 km W Yanahuaya, 2,210 m (N = 28; MVZ 172849 and MVZ 172850).

*andinus.*—Peru: Depto. Arequipa, 2 km W Sumbay, 4,200 m (N = 23; MVZ 174062 and MVZ 174063); Loma Huanucancha, 28 km by road N Arequipa, 3,850 m (N = 1; no DNA).

*boliviensis.*—Peru: Depto. Puno, 12 km S Santa Rosa, 3,950 m (N = 19; MVZ 171607 and MVZ 171608).

*juninensis.*—Peru: Depto. Junin, 22 km N La Oroya, 4,040 m (N = 7; MVZ 173038 and MVZ 173039).

*ko fordii.*—Peru: Depto. Puno, Agualani, 9 km N Limbani, 2,840 m (N = 11; MVZ 171665 and MVZ 171666).

*mollis.*—Peru: Depto. Junin, 16 km NNE Palca, 2,540 m (N = 16; MVZ 173057 and MVZ 173058).

*puer.*—Peru: Depto. Puno, 12 km S Santa Rosa, 3,950 m (N = 6; MVZ 171612).

*subfuscus.*—Peru: Depto. Arequipa, 15 km S Callalli, 4,150 m (N = 87; MVZ 174109 and MVZ 174110); Depto. Cusco, 26 km NW Ollantaytambo, 3,700 m (N = 12; MVZ 174229 and MVZ 174230); Depto. Puno, 6.5 km SW Ollachea, 3,350 m (N = 35; MVZ 172969 and MVZ 172970).

*torques.*—Peru: Depto. Cusco; 3 km E Amaybamba, 2,200 m (N = 13; MVZ 174053 and MVZ 174054), and 32 km NE Paucartambo, 3,140 m (N = 34; MVZ 171720 and MVZ 171721).
Results and Discussion

Sequence Variation in mtDNA

Sequence variation in cyt b in akodontines is illustrated in the autoradiogram in figure 1. This includes examples of variation within a local population, among populations within a species, and between different species and genera. Sequence data for 401 bp of cyt b (fig. 2) were obtained for 20 populations representing nine species of Akodon (subgenus Akodon) and one species in each of three related taxa currently
considered as separate genera or subgenera, including *Microxus*, *Bolomys*, and *Chroeomys* (see table 1).

Two hundred forty (60%) of the 401 bases were identical in the house mouse (*Mus*) and in all of the *Akodon* group. Eighteen sites distinguished all of the *Akodon* relative to *Mus*, and there were 23 variable sites containing only autapomorphies in *Akodon* populations. This left 120 sites (30%) that were phylogenetically informative for this group of akodontine taxa.

Of the 161 sites that are variable in *Mus* or the *Akodon* complex, 114 (70.8%) are variable third positions of codons, 37 (23.0%) are variable first positions, and 10 (6.2%) involve second-position changes. The pattern of variation at the different codon positions conforms to that found in functional mitochondrial protein-coding genes, in that 85.7% of third-position sites in the 401-bp fragment are variable, 27.6% of first-positions sites are variable, and only 7.5% of second-position sites are variable.

The breakdown for the 120 phylogenetically informative sites within the akodontines is as follows: 94 sites (78.3%) involve differences at the third position of the codon, 21 (17.5%) are first-position differences, and five (4.2%) are variable second positions. Of the 94 informative third-position changes, 31 (33.0%) involve transversions. A slightly different way to characterize the kinds of changes that have taken place within the akodontines is to calculate the number of silent versus replacement transitions and transversions for each of the three codon positions. Pairwise comparisons within the akodontines, with one individual representing each local population, give the totals shown in table 2. Most first- and second-position transitions in the data set result in a change in amino acid, while none of the third-position transitions do.
Similarly, most first- and second-position transversions result in a change in amino acid, whereas most third-position transversions do not.

The increase in sequence difference (percentage of sites that differ) with more distantly related populations/taxa is illustrated in figure 3a. Variation within local populations is <1%. Comparisons of individuals from the same species but from different geographic localities show ≤8% sequence difference. The highest values at this level are from comparisons of *A. aerosus* populations, which differ in karyotype (Paucartambo, 2N = 22; Ollachea and Yanahuaya, 2N = 38; San Ramon, 2N = 40). It is likely that these chromosomal forms of *A. aerosus* in fact represent more than one species. If true, the within-species-level comparisons would range only to a maximum of 5%. The among-species differences are 3%-21%, with the highest values being for comparisons of *A. (Chroeomys) jelskii* and *A. andinus* with all of the other *Akodon*, including *Micrurus and Bolomys* (see also the discussion of phylogenetic relationships below). The final level of comparison is that of *Akodon* and its relatives with the Old World murid rodent genus *Mus*, for which the complete mtDNA sequence is known (Bibb et al. 1981). *Mus* differs from these South American murids at 21%-25% of the bases.

Pairwise comparisons of the number of differences between taxa are separated into transitions and transversions in table 3. Since the second individual in a local population differed by no more than three sites (<1% sequence difference), only one individual was chosen to represent each population. Intrapopulation comparisons are given on the diagonal of table 3. The number of transversions increases with more
Comparisons among different species within the genus *Akodon* have a mean number of transversions of 5.15 (range 0–10). The TS/TV within *Akodon* is 7.70. The mean number of transversions for *Microxus* versus *Akodon* is 4.93 (range two to eight), while TS/TV = 8.00. *Bolomys* versus *Akodon* and *Microxus* yields a mean of 10.56 transversions (range 8–13), with TS/TV = 3.95. *Chroeomys* versus *Akodon*, *Microxus*, and *Bolomys* gives a mean of 30.37 transversions (range 27–34), with TS/TV = 1.49. Finally, comparisons of *Mus* versus the akodontines give a mean of 41.65 transversions (range 37–46), with TS/TV = 1.28.

Transition bias has been noted in vertebrate mtDNA (Brown et al. 1982; Higuchi et al. 1984, 1987). A visual illustration of the effect if one were to weight transversions relative to transitions, to compensate for transition bias, is provided in figure 3b. For this part of the figure, the data from table 3 were weighted so that each transversion counted 10 times as much as a transition, according to the method of Higuchi et al. (1984). This approach results in two very distinct groups at the among-*Akodon*-species level of comparison. Values of sequence difference <44% are for comparisons either
within *Akodon* (sensu stricto) and *Bolomys* or between *A. (Chroeomys) jelskii* and *A. andinus*. Values >77% are for comparisons between *Chroeomys* (including *andinus*) and the rest of the *Akodon*. Thus, this treatment emphasizes the distinctness of *A. (Chroeomys) jelskii* and *A. andinus* when compared with the rest of the *Akodon* plus *Microxus* and *Bolomys*.

**Phylogeny Estimation**

**mtDNA Sequences**

The nucleotide sequences were used to estimate a phylogeny by unweighted parsimony. Five hundred repetitions of DNABOOT were run, resampling characters with replacement. The majority-rule consensus tree based on bootstrapping provides the hypothesis of relationships shown in figure 4. Seven equally parsimonious trees were produced by the DNAPARS procedure. A majority-rule consensus tree (not shown) derived from these trees gave essentially the same pattern as that shown in figure 4. Two species, *A. (Chroeomys) jelskii* and *A. (Akodon) andinus*, fall well outside a group that includes the remaining species of *Akodon* (sensu stricto) as well as the currently recognized genera *Microxus* and *Bolomys*. *Bolomys* falls just outside the cluster of *Akodon* and *Microxus*. Within the *Akodon/Microxus* clade, the grouping together of the three geographic samples of the species *A. subfuscus* is highly stable, occurring in >99% of the bootstrap replicates, as is the joining of the two geographic samples of *Microxus* and of the two samples of *A. aerosus* with the same karyotype (2N = 38).

For deeper levels of relationships within a group, the “noise” caused by multiple hits at silent sites over longer periods of time can be reduced either by focusing on transversions or by analyzing the differences in the amino acids encoded by the sequenced nucleotides. There are 39 sites with phylogenetically informative transversions in the *Akodon* data set. The majority-rule consensus tree based on 500 bootstrap replicates when only transversions are used is shown in figure 5. In this analysis, the clustering of *A. andinus* with *A. (Chroeomys) jelskii* outside of *Bolomys*, *Akodon*, and *Microxus* occurs in >99% of the bootstrap replicates. Twenty-nine of a total of 133 amino acid sites are variable in the *Akodon* data set; 13 of these are phylogenetically informative within the akodontines. The PROTPARS procedure in PHYLIP produced 71 equally parsimonious trees from this data set. In the majority-rule consensus tree (fig. 6), *A. andinus* and *A. (Chroeomys) jelskii* again cluster together outside the rest of the *Akodon* in all 71 trees, and *Microxus* continues to be buried deep within the subgenus *Akodon*.

**Relationships Based on Electrophoretic Data Compared with mtDNA Sequence Data**

Relationships based on an electrophoretic analysis of 26 loci in 13 species of akodontine rodents (Patton et al. 1989) are concordant with the major conclusions regarding phylogenetic relationships within the group of taxa considered here. The currently recognized genus *Microxus*, as represented by the type species *mimus*, falls within a clade otherwise containing only members of the subgenus *Akodon*. *Bolomys*, as represented by the type species *amoenus*, is only slightly differentiated from *Akodon* and *Microxus*. *Akodon (Chroeomys) jelskii* is among the most distinct members of the Akodontini—equally so as are other genera such as *Oxymycterus* and much more so relative to *Akodon* (sensu stricto) than is *Bolomys*, for example. Relationships based on electrophoretic data for a set of samples that match those used in the mtDNA data set are illustrated in a Wagner tree (fig. 7). This tree includes 10 species (12
from Patton et al. (1989), plus nine populations including two species (A. andinus and A. kofordi) presented in the present paper for the first time. From figure 7, A. andinus is clearly most closely related to A. (Chroeomys) jelskii, and these two taxa together form a clade well removed from Akodon (sensu stricto).

The major features of relationships among the genera/subgenera in the akodontine group are identical in both the electrophoretic and mtDNA analyses; Microxus forms part of Akodon (sensu stricto), Bolomys joins just outside the Akodon/Microxus cluster, and A. andinus is aligned with A. (Chroeomys) jelskii in a separate taxon very distinct from the other Akodon (including Microxus and Bolomys). The fine-scale relationships of the terminal taxa within the Akodon clade are not, however, identical in the electrophoretic and mtDNA trees. In both analyses the terminal branches are not well resolved, because of high similarity among many of the taxa. Many of the extant species of Akodon probably evolved quite recently, particularly those at high elevations in the Andes where speciation was likely affected by Pleistocene glacial cycles (Patton et al. 1989; Myers et al. 1990).

The electrophoretic data come from 26 protein loci encoded by nuclear genes, with the usual mode of Mendelian inheritance. The cyt b sequences represent part of one protein coding gene in the mitochondrial genome, which is maternally inherited with no recombination (see reviews in Avise et al. 1987; Harrison 1989). In some cases nuclear-DNA and mtDNA phylogenies are discordant (e.g., see DeSalle and Giddings 1986; for review, see Harrison 1989). However, for the akodontines in the present study, the two different molecular genealogies provide identical views of relationship, both among the organisms at the generic/subgeneric level and, in several cases, at a finer scale for populations within species. The general concordance between nuclear and mitochondrial perspectives of phylogenetic relationships indicates that lineage sorting of mtDNA genomes has not been significant in the diversification of this group of rodents.

**Taxonomic Conclusions**

Our understanding of the phylogenetic relationships among members of the akodontine complex of South American murid rodents is still in its early stages. Nevertheless, it is clear from the data presented above (as well as from the data in Patton et al. 1989) that major adjustments in the current taxonomy of the group will be required as additional sequence and other data become available. Minimally, however, at this time it is apparent that reassessment of the generic/subgeneric status of certain taxa must be made. For one thing, if Bolomys is accorded generic status (as argued strongly in Reig 1987), then Chroeomys (including both jelskii and andinus) must be elevated to a genus, if monophyletic taxa are to be maintained. In the reverse case, the generic status of Microxus cannot continue to be supported, since the type species (mimus) clearly falls within a complex of others, all of which are currently placed within the nominant subgenus Akodon.

**Sequence Availability**

These sequences have been deposited in GenBank under accession numbers M35691–M35716.

**Acknowledgments**

Monica M. Frelow, Cristian Orrego, and Thomas J. White provided advice and assistance with PCR and sequencing. Partial material support was provided by Cetus
Corporation through the auspices of Thomas J. White. We had helpful discussions with Thomas D. Kocher, W. Kelley Thomas, Scott V. Edwards, Francis X. Villablanc, and Enrique P. Lessa. Gary R. Smith assisted with data analysis. Specimens were collected under the auspices of the Ministerio de Agricultura and the Museo Historia Natural “Javier Prado”, Lima. Aid in the field was provided by Carol P. Patton, Robert E. Jones, John E. Cadle, Elena Vivar P., and Daniel Valle B., and logistical support in Peru was provided by B. Anthony Luscombe and Renato Marin. Financial support was provided by the Museum of Vertebrate Zoology and the National Geographic Society.

LITERATURE CITED


WESLEY M. BROWN, review editor

Received February 5, 1990; revision received July 23, 1990

Accepted July 23, 1990
Mitochondrial DNA Polymorphisms in the Two Subspecies of \textit{Drosophila sulfurigaster}: Relationship between Geographic Structure of Population and Nucleotide Diversity\textsuperscript{1}

Koichiro Tamura, Tadashi Aotsuka, and Osamu Kitagawa

Department of Biology, Tokyo Metropolitan University

Recent empirical and theoretical studies on mitochondrial DNA (mtDNA) variation in higher animals have suggested that the extent of mtDNA polymorphism is largely affected by spatial population subdivision. To examine this we studied mtDNA polymorphism in two subspecies of \textit{Drosophila sulfurigaster}: \textit{D. s. albostrigata} and \textit{D. s. bilimbata}. \textit{Drosophila sulfurigaster albostrigata} is mainly distributed on the mainland of Southeast Asia. In contrast, \textit{D. s. bilimbata} forms discontinuous populations on many islands scattered in the Pacific Ocean. Because of the difference in their distribution patterns, the two subspecies are thought to be different in the extent of spatial population subdivision. MtDNA was isolated from $>50$ isofemale strains for each subspecies and were analyzed by eight restriction endonucleases. Nucleotide diversity within a population was higher in \textit{D. s. albostrigata} than in \textit{D. s. bilimbata}. However, haplotype diversity was 1.6 times greater in \textit{D. s. bilimbata} (0.85) than in \textit{D. s. albostrigata} (0.53). The large difference in overall heterogeneity was attributed to the difference in interpopulational nucleotide diversity. For the two subspecies the proportion of interpopulational gene diversity in a subdivided population was calculated to be 0.54 in \textit{D. s. bilimbata} and 0.40 in \textit{D. s. albostrigata}. These observations indicate that spatial population subdivision is a major factor in determining mtDNA polymorphism in these subspecies. The extent of mtDNA divergence between the subspecies was very high. The average nucleotide divergence between them was 7.6\%, which is almost the interspecific level reported for other \textit{Drosophila} species. The cause of the high degree of mtDNA divergence is discussed.

Introduction

Extensive polymorphism of mitochondrial DNA (mtDNA) has been reported in many species of higher animals (Avise and Lansman 1983; Avise 1986). These species also revealed geographic differentiation of mtDNA haplotype frequencies. By contrast, humans (Brown 1980), house mice (Ferris et al. 1983), and some marine organisms (Graves et al. 1984; Avise et al. 1986) have shown a low degree of polymorphism and a low degree of geographic differentiation of haplotype frequencies. At the mtDNA level there are some common features among these less polymorphic and less diversified species: lack of geographic barriers within their range (Avise et al. 1987), high mobility (Ferris et al. 1983), and unusual patterns of life history (Avise et al. 1986). On the basis of these observations and theoretical studies (Avise et al.

\textsuperscript{1} Key words: mitochondrial DNA, restriction-map variation, \textit{Drosophila sulfurigaster}, population subdivision.

Address for correspondence and reprints: Tadashi Aotsuka, Department of Biology, Faculty of Science, Tokyo Metropolitan University, 2-1-1 Fukazawa, Setagaya-ku, Tokyo 158.

© 1991 by The University of Chicago. All rights reserved.
0737-4038/91/0801-0006$02.00

104
Mitochondrial Polymorphisms in *Drosophila sulfurigaster* 105

1984; Takahata and Palumbi 1985), it was argued that the extent of mtDNA polymorphism empirically observed in a species is dependent on its contemporary structure of geographic distribution rather than on its evolutionary history.

Comparison of mtDNA polymorphisms among species which have different patterns of distribution may clarify the effect of spatial population subdivision on mtDNA polymorphism. In such a case, however, we cannot use distantly related species, because genetic conditions that affect intraspecific variability of mtDNA, e.g., the modes of transmission and mutation rate, are apparently different among different kinds of organisms. Moreover, each species has its own life history and demographic peculiarities, which affect the mtDNA diversity (Chapman 1989). Therefore, to substantiate the relationship between mtDNA polymorphism and spatial population subdivision, it is necessary to compare in the same (or closely related) species mtDNA polymorphism among populations which are geographically divided to various extents.

In the present study, we examined mtDNA polymorphism in two subspecies of *Drosophila: D. sulfurigaster albostrigata* and *D. s. bilimbata*. These two subspecies are the members of the *D. nasuta* species subgroup and are genetically very closely related (Takanashi 1983). *Drosophila sulfurigaster albostrigata* is primarily distributed in a continuous fashion over the Southeast Asiatic continent. *Drosophila sulfurigaster bilimbata* forms widely scattered populations on many islands in the Pacific Ocean (Kitagawa et al. 1982). The difference in geographic structure of habitat between these two subspecies may provide a good opportunity to study the effect of spatial population subdivision on mtDNA polymorphism.

**Material and Methods**

**Collection of Flies**

Wild strains of *Drosophila sulfurigaster albostrigata* and *D. s. bilimbata* were collected from 19 localities in Southeast Asia and several islands on the Pacific Ocean (table 1). Since *D. s. albostrigata* is distributed continuously on the mainland of Southeast Asia, we grouped small collection sites into four large regions which were designated as Philippines, Borneo island, Malay peninsula, and Thailand-Burma (referred to as populations in table 1).

**Laboratory Procedure**

mtDNA was isolated from adult flies of an isofemale line. mtDNA was extracted from adult flies (1–5 g) by using the alkaline lysis procedure as described by Tamura and Aotsuka (1988), with some modifications for larger-scale extraction.

Restriction-endonuclease digestions of all mtDNA samples were performed using eight enzymes (*HindIII*, *EcoRI*, *XbaI*, *PstI*, *SacI*, *AvaI*, *HaeIII*, and *MspI*) according to the supplier's instruction (Nippon Gene Co., Ltd.).

Gel electrophoresis was carried out with 0.7%–1.2% horizontal, submarine, agarose slab gel in TAE (40 mM Tris-acetate, 1 mM ethylenediaminetetraacetate) buffer (Maniatis et al. 1982). The gel was stained with ethidium bromide, and the digest profile was visualized by illumination with ultraviolet light.

Restriction-cleavage maps of mtDNA were constructed by analyzing the lengths of single- and double-digested fragments. The length of the fragments was determined on the gel by using *SstI* digests of lambda phage DNA as standards. Sequential double digestion (Maniatis et al. 1982) was also carried out to map the sites for the 4-bp recognition enzymes (*HaeIII* and *MspI*).
Table 1
Geographic Origin and Number of Isofemale Lines of Drosophila sulfurigaster albostrigata and D. s. bilimbata Used

<table>
<thead>
<tr>
<th>Population</th>
<th>Collection Site (year)</th>
<th>No. of Strains</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D. s. albostrigata:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Philippines</td>
<td>Los Banos, Philippines (1979)</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Puerto Princesa, Philippines (1979)</td>
<td>4</td>
</tr>
<tr>
<td>Borneo Island</td>
<td>Sandakan, Sabah, Malaysia (1979)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Miri, Sarawak, Malaysia (1979)</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Kuching, Sarawak, Malaysia (1979)</td>
<td>4</td>
</tr>
<tr>
<td>Malay Peninsula</td>
<td>Singapore (1979)</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Kuala Lumpur, Malaysia (1979)</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Penang, Malaysia (1979)</td>
<td>4</td>
</tr>
<tr>
<td>Thailand-Burma</td>
<td>Nakhon Nayok, Thailand (1979)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Chiangmai, Thailand (1981)</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Rangoon (1982)</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Mandalay, Burma (1982)</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Shillong, India (1981)</td>
<td>4</td>
</tr>
<tr>
<td><strong>D. s. bilimbata:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guam</td>
<td>Guam (1981)</td>
<td>10</td>
</tr>
<tr>
<td>Hawaii</td>
<td>Oahu (1981)</td>
<td>10</td>
</tr>
<tr>
<td>Samoa</td>
<td>Pago Pago, American Samoa (1981)</td>
<td>7</td>
</tr>
<tr>
<td>Vava'u</td>
<td>Vava'u, Tonga (1981)</td>
<td>10</td>
</tr>
<tr>
<td>Tongatapu</td>
<td>Tongatapu, Tonga (1981)</td>
<td>10</td>
</tr>
<tr>
<td>Fiji</td>
<td>Nadi, Fiji (1981)</td>
<td>9</td>
</tr>
</tbody>
</table>

Data Analysis

The sequence divergence in base substitutions per nucleotide site (d) between mtDNA haplotypes was estimated from the proportion of the shared restriction sites by using formula (9) of Nei and Li (1979). Calculations of d for restriction enzymes having six-base or four-base recognition sites were done separately and were averaged according to the total number of nucleotides recognized by both types of enzymes.

The extent of mtDNA polymorphism in subspecies was studied by haplotype diversity (\(h\)) and nucleotide diversity (\(\pi\)) [eq. (7) and (18) of Nei and Tajima (1981), respectively]. The extent of restriction-site difference between populations (\(d_{\text{A}}\)) was measured by Nei and Tajima's equation (24), whereas the proportion of interpopulational gene diversity in a subdivided population (\(G_{st}\)) was measured by Nei's (1987) equation (8.27).

Results

An example of the patterns of mtDNA digestion by HindIII is shown in figure 1. In Drosophila sulfurigaster bilimbata no variation in total mtDNA molecular size was found (lanes 6–10). By a comparison of the mobility of mtDNA digests on gels with markers, the size of mtDNA in this subspecies was estimated to be ~16,200 bp. On the other hand, three length variants were found in D. s. albostrigata, and the estimated lengths of these variants were 16,150, 16,350, and 16,550 bp. Comparison of the restriction maps (mentioned below) for these three length variants with the sequence data of D. yakuba (Clary and Wolstenholme 1985) suggested that the length variations in D. s. albostrigata occurs in the A + T-rich region of mtDNA. This was confirmed by our sequencing analyses (data not shown). Aside from the length vari-
Mitochondrial Polymorphisms in *Drosophila sulfuri*gaster

![Image of gel electrophoresis](image_url)

**FIG. 1.**—*Hind*III digestion profiles of mtDNA that were observed in different strains of *Drosophila sulfuri*gaster *albostrigata* (lanes 1–5) and *D. s. bilimbata* (lanes 6–10). The fragments were electrophoresed on a 0.7% agarose gel in TAE buffer. Lane M, *Styl*I digests of the lambda phage DNA, used as molecular-weight markers.

ation, the mtDNA genome size of the two subspecies of *D. sulfuri*gaster was comparable to that of other *Drosophila* species (Wolstenholme and Clary 1985; DeSalle et al. 1986a; Solignac et al. 1986).

With eight restriction enzymes a total of 35 and 36 restriction cleavage sites were mapped, respectively, in 52 strains of *D. s. albostrigata* and in 56 strains of *D. s. bilimbata*. With respect to the presence or absence of each restriction site, six (A1–A6) and 11 (B1–B11) different mtDNA haplotypes were recognized in *D. s. albostrigata* and *D. s. bilimbata*, respectively (fig. 2). Each haplotype within a subspecies differed from the other haplotypes by one to seven gains or losses of sites in *D. s. albostrigata* and by one to five gains or losses in *D. s. bilimbata*. Between subspecies, mtDNA haplotypes differed by 13–22 gains or losses of sites.

The sequence divergence between mtDNA haplotypes, i.e., *d*, was estimated to be 0.33%–2.70% in *D. s. albostrigata*, 0.31%–1.70% in *D. s. bilimbata*, and 5.33%–10.08% between the two subspecies (table 2). Figure 3 shows the dendrogram of mtDNA haplotypes, which has been constructed from the data in table 2 by UPGMA (Sneath and Sokal 1973).

The geographic distribution of each mtDNA haplotype is shown in table 3. In *D. s. albostrigata* Al was the most common haplotype in the all populations except in the Philippines. The *h* value for the entire population of this subspecies was 0.532.
Fig. 2.—Restriction maps of mtDNA haplotypes observed in *Drosophila sulfurigaster albostrigata* and *D. s. bilimbata*. A = AvaI; E = EcoRI; Ha = HaeIII; Hi = HindIII; M = MspI; P = PstI; S = SacI; X = XbaI. a, *Drosophila sulfurigaster albostrigata*. Except for haplotype A1, only the different sites as compared with haplotype A1 are shown. The sites indicated by an asterisk (*) represent the deletion of the site from A1. b, *Drosophila sulfurigaster bilimbata*. Except for haplotype B1, only the different sites as compared with haplotype B1 are shown.
### Table 2
Estimated Nucleotide Divergence $d(\times 100)$ between *Drosophila sulphurigaster* mtDNA Haplotypes

<table>
<thead>
<tr>
<th></th>
<th>D. S. ALBOSTRIGATA</th>
<th>D. S. BILIMBATA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A2</td>
<td>A3</td>
</tr>
<tr>
<td>A1</td>
<td>0.33</td>
<td>0.68</td>
</tr>
<tr>
<td>A2</td>
<td>0.33</td>
<td>0.70</td>
</tr>
<tr>
<td>A3</td>
<td>1.07</td>
<td>1.82</td>
</tr>
<tr>
<td>A4</td>
<td>1.49</td>
<td>1.15</td>
</tr>
<tr>
<td>A5</td>
<td>1.15</td>
<td>5.97</td>
</tr>
<tr>
<td>A6</td>
<td>5.97</td>
<td>1.06</td>
</tr>
<tr>
<td>B1</td>
<td>0.33</td>
<td>0.67</td>
</tr>
<tr>
<td>B2</td>
<td>0.32</td>
<td>0.66</td>
</tr>
<tr>
<td>B3</td>
<td>0.31</td>
<td>0.33</td>
</tr>
<tr>
<td>B4</td>
<td>0.63</td>
<td>0.97</td>
</tr>
<tr>
<td>B5</td>
<td>0.33</td>
<td>0.70</td>
</tr>
<tr>
<td>B6</td>
<td></td>
<td>1.34</td>
</tr>
<tr>
<td>B7</td>
<td></td>
<td>0.32</td>
</tr>
<tr>
<td>B8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In *D. s. bilimbata*, a different haplotype was the most common in each population—except for Guam and Hawaii, in both of which B1 was most common. The $\hat{h}$ value in *D. s. bilimbata* was 0.849, which was significantly higher than that of *D. s. albostrigata*. The $\pi$ value of *D. s. bilimbata* (0.0063) was also higher than that of *D. s. albostrigata* (0.0041).

The mean number of restriction-site differences within a geographic population ($\hat{v}_X$) and the net $\Delta\pi$ are presented in table 4. The $\hat{v}_X$ values were generally higher in *D. s. albostrigata* [average $\hat{v}_X (\bar{v}_X) = 0.952$] than in *D. s. bilimbata* ($\bar{v}_X = 0.605$), whereas $\Delta\pi$ values were substantially higher in *D. s. bilimbata* [average $\Delta\pi (\bar{\Delta}\pi) = 1.517$] than in *D. s. albostrigata* ($\bar{\Delta}\pi = 0.412$).

**Discussion**

Both $\hat{h}$ and $\pi$ were $\sim 1.5$ times as high in *Drosophila sulfurigaster bilimbata* as in *D. s. albostrigata*. In *D. s. albostrigata* six mtDNA haplotypes were found among 52 strains, 34 (65.4%) of which had haplotype A1, whereas in *D. s. bilimbata* a total of 11 different mtDNA haplotypes were found among 56 strains. Geographic differentiation of mtDNA haplotype frequencies was significant in *D. s. bilimbata* but not in *D. s. albostrigata* (see table 3). The average $\Delta\pi$ was considerably higher in *D. s. bilimbata* (1.52) than in *D. s. albostrigata* (0.41). Conversely, $\hat{v}_X$ was higher in *D. s. albostrigata* (0.95) than in *D. s. bilimbata* (0.61). To evaluate the interpopulational divergence of mtDNA, the proportion of mtDNA variability attributable to population differentiation, i.e., $G_a$ (Nei 1987), was estimated for each subspecies. $G_a$ was considerably higher in *D. s. bilimbata* (0.541) than in *D. s. albostrigata* (0.396). Thus, the large difference in the extent of $\hat{h}$ between the two subspecies can be attributed to
### Table 3
Number of mtDNA Clones in Each Geographic Population of *Drosophila sulfurigaster*

#### A. *D. s. albostrigata*

<table>
<thead>
<tr>
<th>mtDNA Haplotype</th>
<th>POPULATION</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>A6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philippines</td>
<td>Borneo island</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>Malay peninsula</td>
<td></td>
<td>10</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>Thailand-Burma</td>
<td></td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>34</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>52</td>
</tr>
</tbody>
</table>

#### B. *D. s. bilimbata*

<table>
<thead>
<tr>
<th>mtDNA Haplotype</th>
<th>POPULATION</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
<th>B5</th>
<th>B6</th>
<th>B7</th>
<th>B8</th>
<th>B9</th>
<th>B10</th>
<th>B11</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guam</td>
<td>Hawaii</td>
<td>9</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>Samoa</td>
<td>Fiji</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Vava'u</td>
<td>Tongatapu</td>
<td>6</td>
<td>1</td>
<td></td>
<td>2</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Tongatapu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>14</td>
<td>1</td>
<td>2</td>
<td>9</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>13</td>
<td>1</td>
<td>6</td>
<td>56</td>
</tr>
</tbody>
</table>

the difference in the degree of the interpopulational diversity. Moreover, both values of $G_{st}$ for the mtDNA variation are considerably higher than the $G_{st}$ for isozyme variation [0.0534 in *D. s. albostrigata* and 0.2011 in *D. s. bilimbata* (authors' unpublished data) and <0.1 in most organisms (Nei 1987)]. This indicates that spatial population subdivision has an important effect on mtDNA diversity.

Sequence divergence between mtDNA haplotypes in several *Drosophila* species has been estimated by restriction analysis: 1.49%-3.29% in *D. simulans* (Solignac et al. 1986), 0.5%-3.3% in *D. silvestris* (DeSalle et al. 1986b), 0.2%-4.6% in *D. heteroneura* (DeSalle et al. 1986b), and 0.7%-4.1% in *D. albomicans* (Chang et al. 1989). Our estimates for the two subspecies of *D. sulfurigaster* were 0.33%-2.70% in *D. s. albostrigata* and 0.31%-1.80% in *D. s. bilimbata* (table 2). These values were in the range that typically was reported for the sequence divergence of mtDNA haplotypes in *Drosophila*.

The nucleotide sequences of *D. s. albostrigata* and *D. s. bilimbata* differ by an average of 7.59% (range 5.33%-10.08%; table 2 and fig. 3). This magnitude of differentiation is very large and reaches almost the interspecific level observed in the *D. melanogaster* species subgroup (Solignac et al. 1986). This suggests that these two subspecies have been separated for a long evolutionary time, despite the fact that they are not fully reproductively isolated (Takanashi 1983).

It is known that introgressive events cause the sharing of the same mtDNA haplotypes between different species with incomplete isolation mechanisms (Ferris et al. 1983; Powell 1983; Solignac 1986). Although the reproductive isolation between *D. s. albostrigata* and *D. s. bilimbata* is too weak to prevent mtDNA gene flow (Ki-
Mitochondrial Polymorphisms in *Drosophila sulfurigaster* 113

Table 4

<table>
<thead>
<tr>
<th></th>
<th>(i) (on the diagonal) and  (d_A) in <em>Drosophila sulfurigaster</em></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A. <em>D. s. albostrigata</em></td>
</tr>
<tr>
<td></td>
<td>Philippines</td>
</tr>
<tr>
<td>Philippines</td>
<td>0.250</td>
</tr>
<tr>
<td>Borneo island</td>
<td>1.818</td>
</tr>
<tr>
<td>Malay peninsula</td>
<td></td>
</tr>
<tr>
<td>Thailand-Burma</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tagawa et al. 1982; Takanashi 1983), there was no sign of introgressive hybridization between them. On the contrary, the high level of mtDNA divergence observed suggested that spatial or geographical isolation can cause differentiation of mtDNA between taxa even in the absence of complete reproductive isolation.

Acknowledgments

We thank Dr. Masatoshi Nei for helpful discussion. This research was partly supported by the funds of the Overseas Scientific Expedition in 1979 (grant 404149) and 1981 (grant 56041049) of the Ministry of Education, Science and Culture of Japan.

LITERATURE CITED


Masatoshi Nei, reviewing editor

Received February 5, 1990; revision received August 3, 1990

Accepted August 3, 1990
The Relative Rate of DNA Evolution in Primates

Simon Easteal

Human Genetics Group, John Curtin School of Medical Research, The Australian National University

In 73 relative-rate tests involving the sequences of 17 genes between humans and six nonhuman primate taxa, there is only one significant ($P < 0.01$) difference in evolutionary rate—i.e., that between human and Old World–monkey $\beta$-globin genes. No evolutionary rate difference between humans and Old World monkeys is evident from analysis of 18 other genes with a total length of 6 kb. This and the comparison, between humans and other primate taxa, of new extended $\beta$-globin sequences suggest that earlier observations of evolutionary-rate differences between humans and other primates were based on differences that are peculiar to $\beta$-globin and that are not representative of the whole genome, which appears to be evolving at a stochastically uniform rate. This is supported by whole-genome single-copy DNA and mitochondrial DNA comparisons, neither of which shows any evidence of evolutionary-rate variation among primate taxa. Uniformity in the evolutionary rate of the DNA of primate and other mammalian taxa is inconsistent with current mammalian fossil-record interpretation. Either there has been a general slowing down in rate across lineages or the fossil record has been misinterpreted.

Introduction

DNA and protein differences between humans and other apes are less than those expected for the accepted mammalian divergence times when a constant rate of molecular evolution is assumed. Two alternative explanations have been proposed. Either, as first suggested by Goodman (1961), the rate of molecular evolution is slower in primates—and in hominoids in particular—than in other mammals, or, as originally argued by Sarich and Wilson (1967), the fossil record has been misinterpreted in deriving divergence times. The former explanation has been the more widely accepted in recent years (Goodman 1985; Britten 1986; Li and Tanimura 1987; Li et al. 1987; Sibley and Ahlquist 1987), although the latter has retained some support (Hasegawa et al. 1987; Wilson et al. 1987; Easteal 1988, 1990).

The issue can only be resolved by testing for rate variation in a way that does not depend on fossil-record interpretation. One approach is the relative-rate test of Wu and Li (1985), in which the evolutionary rates of two species are compared relative to a third, more distantly related reference species. With this approach some knowledge of phylogeny is necessary because a reference species must be chosen that is the most distantly related species of the three being compared. Beyond this, however, no assumptions need be made about species divergence times.

The relative-rate approach can readily be applied to primates, among which the branching order of strepsirhines, New World monkeys, Old World monkeys, and then
the apes branching in the order of gibbons, orangutans, and finally the African apes is now well established.

The branching order of the African apes remains controversial, but a branching of gorillas before humans and chimpanzees is now indicated by whole-genome, non-repetitive DNA-DNA hybridization studies (Sibley and Ahlquist 1984, 1987; Felsenstein 1987; Caccone and Powell 1989), two-dimensional protein electrophoresis (Goldman et al. 1987), and DNA sequence data (Fitch et al. 1988; Holmquist et al. 1988; Miyamoto et al. 1988; Goodman et al. 1989; Koop et al. 1989; Williams and Goodman 1989).

In the present paper I apply the principle of the relative-rate test to nuclear and mitochondrial DNA sequence differences among primate taxa and also test for evidence of rate variation in the whole-genome DNA-DNA hybridization data. In so doing I demonstrate that, with two exceptions, there is among primate lineages no evidence of variation in rate of DNA evolution—and thus no evidence of a general slowing down of rate in the lineage leading to humans.

Material and Methods

DNA sequences were obtained either from the EMBL data base (1989, release 19) or directly from the published literature. Formal relative-rate tests were performed by the method of Wu and Li (1985). Alignment of noncoding sequences was by an iterative multiway procedure using the computer program ALIGN supplied by D. Smith (Australian National University), except where published multiway sequence alignments were already available. Rates of synonymous and nonsynonymous sites were estimated by the method of Li et al. (1985) by using a computer program supplied by W.-H. Li (University of Texas). Correction for multiple substitutions in noncoding sequences was made by assuming a Poisson distribution (Jukes and Cantor 1969).

Evolutionary-rate variation for whole-genome single-copy DNA was assessed using the F-ratio test (Felsenstein 1984). This test compares the sum of squares (SS) of the phylogenetic trees with the least-squares topology constructed with and without the constraint that branch lengths are equal. Least-squares topologies were obtained by using the algorithms FITCH (branch lengths unconstrained) and KITCH (branch lengths constrained) from the phylogenetic reconstruction package PHYLIP (J. Felsenstein, University of Washington).

Results

Nuclear DNA Sequences

Table 1 shows the results of relative-rate tests for both the sequences analyzed by Li et al. (1987) and for strepsirhine sequences and other sequences, including the expanded region of the \( \gamma \)-globin gene (Fitch et al. 1988; Miyamoto et al. 1988; Goodman et al. 1989), published since Li et al.'s analysis. Additional genes were included in the analysis only if a suitable reference sequence—i.e., an orthologue of a more distantly related primate species in human-ape comparisons or an orthologue from another order of eutherian mammal in human—Old World—monkey comparisons—was available. For alcohol dehydrogenase a closely related human paralogue was used. The involucrin gene sequences (Djian and Green 1989a, 1989b; Teumer and Green 1989; Tseng and Green 1989) were not included because of perceived alignment problems resulting from the repetitive nature of the gene sequence. Minor discrepancies between the results in table 1 and the results of comparisons of the same genes between the same taxa reported by Li et al. (1987) arise either from differences
in reference-species selection or sequence alignment procedures or from expansion of the length of sequence or inclusion of sequence gaps in the analysis. In only one comparison is such a discrepancy significant. Li et al. (1987) reported that between humans and New World monkeys there was a significant difference in the rate of \( \gamma \)-globin gene evolution. In the present analysis, which is based directly on the results reported by Fitch et al. (1988), no such difference is apparent.

A total of 73 comparisons involving 17 genes and six nonhuman primate taxa were made. In only one of these comparisons is there a significant \((P < 0.01)\) rate difference between humans and the other compared species. This comparison is that between human and Old World–monkey \( \gamma \)-globins. This comparison involves nearly 8 kb of sequence, in addition to the 2 kb for which Li et al. (1987) also reported a significant \((P < 0.01)\) difference in rate. For the orangutan, gorilla, and chimpanzee \( \gamma \)-globin comparisons, the same 8 kb of additional sequence is also available (Miyamoto et al. 1988). In the comparisons of these extended sequences the differences in rate are all substantially less than those obtained from the shorter sequences (Li et al. 1987). The orangutan and human rates are now identical. In the human-gorilla \( \gamma \)-globin comparison, where no other additional data were included, the difference for the \( \gamma \)-globin comparison changes the result for the comparison of all genes combined. Where previously there appeared to be a slower rate of evolution in the human lineage, significant at the 0.01 level, there is now no evidence of any overall rate difference between humans and gorillas.

In the rate difference among apes the reduction associated with the increased length of \( \gamma \)-globin sequence suggests that the original observation of significant rate differences may be due to sampling error and that the difference is confined to the 2-kb region originally analyzed. For the human—Old World–monkey comparison, when the \( \gamma \)-globin sequence, which comprises \( \sim 63\% \) of the compared nucleotides, is removed, although the value of \( K_{13} - K_{23} \) is still positive there is no significant difference in rate for the remaining 18 genes, which have a total length of 6 kb \((K_{12} = 8.0; K_{13} - K_{23} = 0.8 \pm 0.5)\).

Overall \( K_{13} - K_{23} \) has a positive value, indicating a slower rate of evolution in humans in only 52% of all comparisons. In the remaining 48% of comparisons \( K_{13} - K_{23} \) has either a negative or zero value. There is thus no overall indication of a slowdown in evolutionary rate in the human lineage.

### DNA Thermostability Comparisons

Nuclear nonrepetitive-DNA thermostability data are consistent with the nuclear-DNA sequence data in showing an absence of rate variation among lineages. Table 2 shows \( \Delta T_{50H} \) (Sibley and Ahlquist 1987) and \( \Delta T_{50R} \) (Benveniste 1985) values above the diagonal and \( \Delta T_m \) values (Caccone and Powell 1989) below the diagonal.

For the \( \Delta T_{50H} \) data Felsenstein (1987) found no evidence of rate variation among lineages. Caccone and Powell (1989) noted that this appears also to be the case for the \( \Delta T_m \) data. Application of the “F-ratio test” (Felsenstein 1984) to the \( \Delta T_m \) data confirms this to be the case \((SS \text{ FITCH} = 0.307; SS \text{ KITCH} = 0.392; F \text{-ratio} = 0.595; F_{0.05; 12,221} = 2.4)\).

The more complete data sets of Sibley and Ahlquist (1987) and Caccone and Powell (1989) are restricted to apes and Old World monkeys. The \( \Delta T_m \) values obtained by Benveniste (1985) for strepsirhine—New World–monkey comparisons, although less complete, show no evidence of rate variation between human and either Old World–monkey or New World–monkey lineages. Old World monkeys and humans
### Table 1
Differences in the Numbers of Substitutions/100 Nucleotides, between Human and Other Primate Genes

<table>
<thead>
<tr>
<th>Gene (N°)</th>
<th>LEMUR</th>
<th>NEW WORLD MONKEY</th>
<th>OLD WORLD MONKEY</th>
<th>PONGO</th>
<th>GORILLA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K_{12}$</td>
<td>$K_{13}-K_{23}$</td>
<td>$K_{12}$</td>
<td>$K_{13}-K_{23}$</td>
<td>$K_{12}$</td>
</tr>
<tr>
<td>Pseudogene</td>
<td>28.9</td>
<td>6.0±2.9</td>
<td>11.1</td>
<td>2.5±1.6</td>
<td>7.9</td>
</tr>
<tr>
<td>Exon silent sites:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>α-Globin (106)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β-Globin (114)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>δ-Globin (105)</td>
<td>37.8</td>
<td>-3.0±7.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ε-Globin (100)</td>
<td>28.5</td>
<td>-11.2±8.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insulin (82)</td>
<td>33.2</td>
<td>16.9±11.0</td>
<td>12.3</td>
<td>1.8±5.1</td>
<td>3.1</td>
</tr>
<tr>
<td>TGF (276)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADH (260)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ig (309)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apo A1 (158)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apo E (225)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPO (145)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pep (263)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a1 AT (140)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POMC (180)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Introns:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>α-Globin (237)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β-Globin (190)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>δ-Globin (655/960)</td>
<td>29.7</td>
<td>-5.0±3.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ε-Globin (863)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ε-Globin (941)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Values are given as mean±SD.
<table>
<thead>
<tr>
<th>Gene</th>
<th>K₁₂</th>
<th>K₁₃</th>
<th>K₂₃</th>
<th>K₁∞</th>
<th>K₂∞</th>
<th>K₃∞</th>
<th>K₁∞</th>
<th>K₂∞</th>
<th>K₃∞</th>
</tr>
</thead>
<tbody>
<tr>
<td>γ-Globin</td>
<td>33.6</td>
<td>-1.4±2.9</td>
<td>13.5</td>
<td>2.2±1.6</td>
<td>3.2</td>
<td>-0.4±0.6</td>
<td>3.0</td>
<td>0.7±0.6</td>
<td>2.6</td>
</tr>
<tr>
<td>Insulin</td>
<td>18.8</td>
<td>-6.7±3.8</td>
<td>2.9</td>
<td>1.6</td>
<td>0.6</td>
<td>20.6</td>
<td>0.6</td>
<td>2.6</td>
<td>-0.1±0.6</td>
</tr>
<tr>
<td>Ige</td>
<td>1.4</td>
<td>1.6</td>
<td>0.7</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Flanking sequences:

<table>
<thead>
<tr>
<th>Gene</th>
<th>K₁₂</th>
<th>K₁₃</th>
<th>K₂₃</th>
<th>K₁∞</th>
<th>K₂∞</th>
<th>K₃∞</th>
<th>K₁∞</th>
<th>K₂∞</th>
<th>K₃∞</th>
</tr>
</thead>
<tbody>
<tr>
<td>α-Globin</td>
<td>5.0</td>
<td>2.4±2.3</td>
<td>4.0</td>
<td>0.5±1.5</td>
<td>3.3</td>
<td>-0.8±1.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β-Globin</td>
<td>10.2</td>
<td>3.4±2.5</td>
<td>7.9</td>
<td>-1.7±1.5</td>
<td>0.6</td>
<td>-0.6±0.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ε-Globin</td>
<td>19.4</td>
<td>0±3.4</td>
<td>9.1</td>
<td>1.9±2.6</td>
<td>6.3</td>
<td>1.0±1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>γ-Globin</td>
<td>24.1</td>
<td>-1.1±2.7</td>
<td>2.9</td>
<td>0.2±0.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insulin</td>
<td>28.3</td>
<td>-5.6±4.0</td>
<td>14.2</td>
<td>-1.7±2.6</td>
<td>6.1</td>
<td>-0.1±1.4</td>
<td>1.3</td>
<td>0.7±0.7</td>
<td>0.6</td>
</tr>
<tr>
<td>TGF</td>
<td>18.6</td>
<td>3.9±2.5</td>
<td>8.1</td>
<td>1.3±2.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POMC</td>
<td>230</td>
<td>15.6</td>
<td>4.7±3.9</td>
<td>2.7</td>
<td>-0.4±0.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ige</td>
<td>539</td>
<td>2.7</td>
<td>4.7±3.9</td>
<td>1.6</td>
<td>0.1±0.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Total** 27.1 -1.3±1.1 | 13.0 | 1.4±0.8 | 7.9 | 1.2±0.3** | 3.6 | 0.0±0.2 | 1.7 | 0.1±0.1 | 1.6 | 0.1±0.1**

**NOTE.**—The relative-rate test involves a comparison (K₁₂) of the rate of substitution (K) between sequence 1 and a reference sequence (3) and between sequence 2 and the reference sequence. In all comparisons the human sequence is sequence 2; thus negative values of K₁₂ indicate a relatively faster rate of substitution in humans. The reference genes were those used by Li et al. (1987), except for the following: rabbit for all lemur comparisons; human β-globin/rabbit β-globin for New World–monkey β-globin; lemur for New World–monkey γ-globin; dog/mouse for New World–monkey insulin; rabbit for Old World–monkey ε- and δ-globins; lemur for Old World–monkey ε-globin; New World monkey for Old World–monkey ε-globin and insulin; mouse for Old World–monkey TGF (transforming growth factor); Old World monkey for orangutan α- and δ-globins; lemur for orangutan ε-globin; New World monkey for orangutan γ-globin; rabbit for Old World–monkey Apo E (apolipoprotein E); human γ ADH (alcohol dehydrogenase) for New World–monkey δ ADH, and pig for Old World–monkey PEP (pepsinogen). In all cases these are the most closely related available sequences. General data sources were GenBank (release 48), EMBL (release 12), Wu and Li (1985), Li and Tanimura (1987), and Eastal (1988). For specific genes, data sources were as follows: Old World–monkey and orangutan α- and β-globins, Shaw et al. (1987); New World–monkey ε-globin, Spritz and Giebel (1988); Old World–monkey insulin, Seino et al. (1987); Old World–monkey TGF, Sharples et al. (1987); human δ-globin, Hsu et al. (1988); Old World–monkey POMC, Patel et al. (1988); human POMC, Takahashi et al. (1981); cow POMC, Nakatsui et al. (1979); human PEP A, Sogawa et al. (1983); Old World–monkey PEP A, Evers et al. (1988); pig PEP A, Tsukagoshi et al. (1988); human ADH, Ikuta et al. (1986); Old World–monkey ADH, Trezise et al. (1989); human Apo E, Breslow et al. (1982); Old World–monkey Apo E, Marotti et al. (1989); and rabbit Apo E, Hao et al. (1987). The results for the gorilla and chimpanzee comparisons, except for γ-globin, were obtained directly from Li et al. (1987). Estimates of the evolutionary rates of the γ-globin genes were obtained from Fitch et al. (1988), Miyamoto et al. (1988), and Goodman et al. (1989). The approximate lengths of the γ-globin sequences compared were as follows: human–chimpanzee, 10,150; human–gorilla, 10,080; human–orangutan, 9,927; human–Old World monkey, 9,417; human–New World monkey, 1,827; and human–strepsirhine, 724.

* Approximate number of sites compared.

** P < 0.01.
### Table 2
Thermostability Differences between Primate Taxa

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Homo</th>
<th>Pan</th>
<th>Gorilla</th>
<th>Pongo</th>
<th>Hylobates</th>
<th>Old World Monkey</th>
<th>New World Monkey</th>
<th>Strepsirhine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo</td>
<td>1.63</td>
<td>2.27</td>
<td>3.60</td>
<td>4.76</td>
<td>7.34</td>
<td>13.1</td>
<td>24.1</td>
<td></td>
</tr>
<tr>
<td>Pan</td>
<td>1.59</td>
<td>2.23</td>
<td>3.57</td>
<td>4.83</td>
<td>7.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gorilla</td>
<td>2.50</td>
<td>2.55</td>
<td>3.55</td>
<td>4.69</td>
<td>7.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pongo</td>
<td>3.49</td>
<td>3.52</td>
<td>3.57</td>
<td>4.83</td>
<td>7.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hylobates</td>
<td>5.04</td>
<td>4.66</td>
<td>5.15</td>
<td>4.83</td>
<td>7.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old World Monkey</td>
<td>6.78</td>
<td>7.01</td>
<td>7.12</td>
<td>7.33</td>
<td>6.98</td>
<td>13.1</td>
<td>24.5</td>
<td></td>
</tr>
<tr>
<td>New World Monkey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24.0</td>
</tr>
</tbody>
</table>

**NOTE.**—Except for those involving New World monkeys and strepsirhine, which are from Benveniste (1985) and have differences estimated as $\Delta T_{mR}$, comparisons above the diagonal are from table 5 of Sibley and Ahlquist (1987) and have differences estimated as $T_{mH}$. $T_{mH}$ values in comparisons with Pan are weighted average values for $P$. paniscus and $P$. troglodytes. $\Delta T_{mR}$ values are average values for humans compared with six species of New World monkeys and with two species of strepsirhine and for one species of New World monkey compared with two species of Old World monkey. Comparisons below the diagonal are the $\Delta T_{m}$ values from Caccone and Powell (1989), with average values for $P$. paniscus and $P$. troglodytes and for $H$. lar and $H$. syndactylus.

The values for these three taxa (24.5, 24.1, and 24.0, respectively) are extremely similar when compared with those for prosimians, with the human value being intermediate between the other two.

**Mitochondrial DNA**

The sequence of a 0.9-kb segment of the mitochondrial genome containing three tRNA genes and parts of two protein coding genes has been determined for 12 primate species (Brown et al. 1982; Hasegawa et al. 1987; Hayasaka et al. 1988). Because of the relatively high rate of transition substitutions in mtDNA, it is necessary to separate transitions from transversions in estimating relative rates of evolution (Hasegawa et al. 1987). If this is not done, the saturating effect of the more rapidly occurring transitions might mask among-lineages rate differences that would be evident for transversions. The among-taxa numbers of transversion substitutions, derived from the combined data of Hasagawa et al. (1987) and Hayasaka et al. (1988), are shown in table 3. There is no indication of any rate differences between any of the lineages; the values down each column of the table are extremely similar. In only one case (Homo and Pan compared with Gorilla) is the number of transitions in the human comparison the lowest. In the comparison of Homo, Pan, and Gorilla with Pongo the human rate is the highest, and in all other comparisons the human rate is within the range of those for other lineages.

Comparison of the human and Old World–monkey mitochondrial cytochrome C oxidase subunit II genes (681 bases) and the cow gene (Ramharack and Deeley 1987) shows a nonsignificantly higher number of silent transversion substitutions in the Old World–monkey comparison (58) than in the human comparison (55)—but the same number (39) of replacement-site transversions. These two regions of mtDNA are thus consistent with the nuclear DNA in showing no evidence of evolutionary-rate variation among primate lineages.
Table 3
Numbers of Transversion-Type Differences between Mitochondrial DNA Sequences of Primate Taxa

<table>
<thead>
<tr>
<th></th>
<th>Pan</th>
<th>Gorilla</th>
<th>Pongo</th>
<th>Hylobates</th>
<th>Old World Monkey</th>
<th>New World Monkey</th>
<th>Strepsirhine* and Tarsiform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo</td>
<td>0.55</td>
<td>0.90</td>
<td>3.92</td>
<td>5.03</td>
<td>8.30</td>
<td>10.97</td>
<td>15.51</td>
</tr>
<tr>
<td>Pan</td>
<td>1.01</td>
<td>3.80</td>
<td>4.91</td>
<td>8.18</td>
<td>11.07</td>
<td>15.52</td>
<td></td>
</tr>
<tr>
<td>Gorilla</td>
<td>3.67</td>
<td>5.03</td>
<td>8.07</td>
<td>12.40</td>
<td>14.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pongo</td>
<td></td>
<td>5.83</td>
<td>8.47</td>
<td>14.98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hylobates</td>
<td></td>
<td></td>
<td>8.63</td>
<td>11.66</td>
<td>14.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old World Monkey</td>
<td></td>
<td></td>
<td></td>
<td>11.44</td>
<td>14.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New World Monkey</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15.89</td>
</tr>
</tbody>
</table>

NOTE.—Data are from Brown et al. (1982) and Hayasaka et al. (1988).
* Values are the average values for lemurs and tarsiers.
b Values for comparisons involving Old World monkeys are the average for four species of Maccaca.

Discussion

The results of the present study show that, when no assumption other than their relative order is made about species divergence times, the rates of DNA evolution between humans and other primates appear generally to be stochastically uniform. Since the view that there is a slowing down of the rate of molecular evolution in the human lineage has become so widely accepted, it is necessary to consider other recent studies to examine how this contrary conclusion has been reached.

Some investigators, such as Britten (1986) and Goodman et al. (1982, 1983), have arrived at the conclusion that primate evolutionary rates are variable by making species-divergence-time assumptions on the basis of fossil-record interpretation. In studies of this kind the validity of the fossil-record interpretation is not being tested; it is assumed, and the validity of a finding of rate variation is only as good as this assumption.

There is ample reason for thinking that the mammalian fossil record is not always correctly interpreted. Thus, for example, the presumed polychotomous divergence of eutherian orders is inconsistent with their hierarchical phylogeny indicated by DNA sequence comparisons (Fasteal 1988, and accepted); and, more pertinent, the taxonomic status of the fossils Sivapithecus and Aegyptopithecus has recently been revised, with a resulting dramatic change in the estimated time of divergence of humans from the other African apes (Wilson et al. 1987).

More interesting than fossil record–based studies have been those that have used the relative-rate approach (Li and Tanimura 1987; Li et al. 1987). Li et al. (1987), in an analysis of the relative evolutionary rates of the sequences of 11 primate genes, found that one, the \( \psi \eta \)-globin gene and its flanking sequences, has evolved more slowly in humans than in other primates. In 10 other nuclear genes no significant differences were observed. However, the human rate was (nonsignificantly) lower than that of the compared species in 28 of the 36 comparisons. In addition, in comparisons of humans with gorillas, Old World monkeys, and New World monkeys, the rate differences were significant when all genes (including \( \psi \eta \)-globin, which constituted \( \sim 60\% \) of the compared nucleotides) were combined. The interpretation that these results demonstrate a slower rate of molecular evolution in humans than in other primates depends very much on the difference observed in a single gene, \( \psi \eta \)-globin.
With one exception, the apparent human slowdown is not evident in the present analysis for comparisons involving more extensive sequence data for the region surrounding \( \psi \)-globin (Fitch et al. 1988; Miyamoto et al. 1988; Goodman et al. 1989) or in a total of 68 other sequence comparisons.

Hasegawa et al. (1987) demonstrated by relative-rate tests that the rate of evolution of a segment of mitochondrial DNA does not vary among the human, chimpanzee, gorilla, orangutan, gibbon, and cow lineages. Hayasaka et al. (1988) analyzed, in addition to these, the homologous mitochondrial DNA sequences from several Old World monkeys, a New World monkey, and two strepsirhines. They showed by relative-rate test that there is no rate variation among the taxa; however, they concluded that the rate had slowed down in the human lineage. This anomalous conclusion was based on consideration of fossil record–derived divergence times. Analysis of the entire data set presented here, as well as of the cytochrome C oxidase subunit II gene, shows that there is no evidence of rate variation in the evolution of mitochondrial DNA.

This result based on analysis of transversion substitutions evident from DNA sequence data would appear to contradict the earlier finding of Templeton (1983a, 1983b). From analysis of restriction-map data, he concluded that the rate of mitochondrial DNA evolution has been slower in the human lineage than in the combined gorilla-chimpanzee lineage. Nei and Tajima (1985), however, showed that assumptions underlying Templeton’s test for rate variation were invalid and that the molecular clock hypothesis could not be rejected by the data he analyzed. Furthermore, Templeton’s analysis was based on a phylogeny that placed chimpanzees and gorillas in a monophyletic group relative to humans, which in the light of the analysis of nucleotide sequence data now seems highly improbable.

Sibley and Ahlquist’s (1984) study of DNA-DNA hybridization is unusual in that it lcd to the conclusion that the rate of evolution based on fossil record–derived divergence times did not vary among lineages. The reason for this is that different divergence times were assumed. Specifically, an orangutan divergence of 16 Mya was assumed, based on the assumption that \textit{Sivapithecus} is uniquely ancestral to orangutans. They support their rate-constancy argument by consideration of the previously estimated evolutionary rates of bird DNA, which were assumed to be the same as the presumed primate rates. However, aside from the fact, recognized by Sibley and Ahlquist (1987), that the bird and mammal rates may be different, the way in which the bird rates were obtained has been questioned (Hounde 1986, 1987).

In the analysis of an expanded data set, Sibley and Ahlquist (1987) conclude that rate does vary among lineages in proportion to age at first breeding. This conclusion, however, is not arrived at through any formal analysis of the data, and Felsenstein (1987) was unable to detect any rate variation among lineages by analysis of the same data set. Caccone and Powell (1989) found no evidence of rate variation in their DNA-DNA hybridization study, and that result is confirmed here.

The one exception to the apparent overall pattern of rate uniformity presented here is the \( \psi \)-globin gene, which has evolved relatively faster in the Old World-monkey lineage than in the human lineage. This rate difference was observed for a 2-kb sequence by Li et al. (1987). It is still apparent for the 10-kb sequence now available, indicating that the difference is real and not due to sampling error. The demonstration of a real difference for the \( \psi \)-globin region does not, however, mean that a difference exists for the genome generally. There are two reasons for thinking that it does not.

First, there is no evidence of any rate difference between the human and Old World–monkey lineages in the comparison of 6 kb of sequence from 18 genes other
than the Ψη-globin gene. Although this is a smaller number of nucleotides than that
compared at the Ψη-globin gene, it is three times as many nucleotides as were originally
sequenced at the Ψη-globin gene and which showed a significant rate difference. If
that rate difference occurred generally throughout the genome, it would be apparent
from comparison of 6 kb of sequence. A sample of 6 kb from 18 genes is a better
indicator of overall evolutionary rate than is a sample of 10 kb for one gene. Second,
the ΔT_m,R values (Benveniste 1985) for the human and Old World-monkey lineages
compared with New World monkeys are identical. While further sequence data are
needed to resolve the issue conclusively, it appears from the data presently available
that Ψη-globin is unusual in showing a rate difference and that generally there is rate
uniformity between the two lineages. It demonstrates the need for caution in inter-
preting results obtained from the comparison of sequences in the region of a sin-
gle gene.

A rate difference has also been demonstrated between two other nonhuman pri-
mate lineages. The noncoding sequences of the ε-, γ-, and Ψη-globin genes have evolved
faster in the galago lineage than in the lemur lineage (Koop et al. 1989). Since this
rate difference has been observed at three separate genes and is consistent with the
DNA-DNA hybridization results (Bonner et al. 1980), it would appear to be a general
property of the single-copy genome. It demonstrates that, even if they are not the
general rule, evolutionary-rate differences between lineages do exist.

Two important conclusions can be drawn from the finding that molecular evo-
lutionary rates are generally uniform both among primates as shown here and between
primates and other mammalian orders (Easteal 1988, 1990). First, the favored ex-
planation for the presumed slower evolutionary rate in humans is that the rate of
molecular evolution is dependent on cell generation time, which is relatively greater
in the human lineage (Goodman 1985; Li and Tanimura 1987; Li et al. 1987). The
absence of lineage-specific differences in evolutionary rate means that the rate of DNA
evolution is not cell-generation-time dependent. As pointed out by Sarich and Wilson
(1973), this would suggest that mutations arise predominantly by processes that are
independent of DNA replication.

Second, since the discrepancy between the observed and expected degree of nu-
cleotide change between apes and other primates is not explained by substitution-rate
differences, it follows that either there has been a general slowing down of molecular
evolutionary rate in the different mammalian lineages during the Cenozoic, as proposed
by Goodman (1985) and Gingerich (1986), or the divergence times of at least some
primate and/or other mammalian taxa are different from those currently proposed;
it also follows that the fossil record has been misinterpreted in some way.

The first possibility cannot be ruled out; however, it is difficult to understand
how such a slowdown could have occurred. A number of factors have been identified
that could cause variation in molecular evolutionary rate; these include natural selec-
tion, variation in cell-generation time (Wu and Li 1985; Li et al. 1987), varying
efficiency of DNA-repair mechanisms (Britten 1986), and differential exposure to
environmental mutagens (Goodman 1985). It seems highly unlikely, however, that
any of these could account for a gradual DNA-evolutionary-rate slowdown that affected
all lineages equally. Goodman et al. (1975) and Gingerich (1986) have suggested that
natural selection might do this with respect to particular genes. An interlineage slow-
down, to be consistent with both the relative-rate analysis presented here and that of
genes in different mammalian orders (Easteal 1988, 1990), must have affected many
genes as well as noncoding sequences and mitochondrial DNA. Natural selection could not account for such a generalized effect.

Cell-generation time varies among lineages and has been put forward as an explanation for apparent interlineage rate variation. It could not account for a uniform rate change among lineages. Variation in the efficiency of DNA-repair mechanisms also has been put forward to explain apparent evolutionary-rate differences among lineages. It is unlikely that the enzymes involved in DNA repair have independently evolved to be more efficient to the same degree in diverse lineages. Similarly, it is difficult to conceive of an environmental mutagen whose effects could have been uniform on organisms with different life histories and living in different environments and in different parts of the world. The effects of a reduction in incident ultraviolet radiation, for example, would vary with latitude, habitat, morphology, and reproductive behavior and physiology.

Although a uniform interlineage slowdown might be hard to envisage, it may nevertheless have occurred. Distinguishing between it and fossil-record misinterpretation will require more extensive investigation of evolutionary rates in different taxonomic groups by using a combination of the relative-rate approach and fossil record-derived divergence times.

Acknowledgments

I am grateful to J. A. Cavanaugh, P. W. Easteal, M. Goodman, and S. W. Serjeantson for comments and discussion. I also thank D. Smith and W.-H. Li for providing computer programs.

LITERATURE CITED


WESLEY M. BROWN, reviewing editor

Received December 14, 1989; revision received July 7, 1990

Accepted July 24, 1990
Methods for Inferring Phylogenies from Nucleic Acid Sequence Data by Using Maximum Likelihood and Linear Invariants

W. C. Navidi, G. A. Churchill, and A. von Haeseler

Department of Mathematics, University of Southern California

Likelihood methods and methods using invariants are procedures for inferring the evolutionary relationships among species through statistical analysis of nucleic acid sequences. A likelihood-ratio test may be used to determine the feasibility of any tree for which the maximum likelihood can be computed. The method of linear invariants described by Cavender, which includes Lake’s method of evolutionary parsimony as a special case, is essentially a form of the likelihood-ratio method. In the case of a small number of species (four or five), these methods may be used to find a confidence set for the correct tree. An exact version of Lake’s asymptotic χ² test has been mentioned by Holmquist et al. Under very general assumptions, a one-sided exact test is appropriate, which greatly increases power.

Introduction

Likelihood methods have long been used in problems of phylogenetic inference, while methods based on linear invariants are of recent origin. Both methods are typically used to find a single tree which best fits some nucleic acid sequence data. We discuss some examples of this type of procedure. We also describe some confidence-set procedures, which are designed to yield collections of plausible trees rather than single estimates.

The use of maximum likelihood in phylogenetic inference is well established. Two papers giving excellent reviews of the subject are by Felsenstein (1983, 1988). Many models have been proposed for generating likelihood functions. Felsenstein (1981) proposed a model with one rate parameter per branch. Among the virtues of this model is that it is easy to maximize the likelihood for trees containing a fairly large number of species. We give below a numerical example using a version of this model. Cavender and Felsenstein (1987) and Barry and Hartigan (1987) describe a very general model with 12 parameters per branch. This enables maximum-likelihood methods to be used with relatively few assumptions. Most of the more restrictive models in common use can be obtained by placing constraints on this model. The general model discussed below is essentially this one, and the likelihood function we use is essentially the one described by Barry and Hartigan (1987) in their “maximum average likelihood” approach.

The first use of linear invariants was due to Lake (1987), who devised a procedure known as “evolutionary parsimony.” One appeal of linear-invariant methods is that the assumption that the evolutionary process is identical in every site is not needed. Instead, some constraints are put on the parameters of the model.

1. Key words: phylogenetic inference, invariants, likelihood, evolutionary parsimony.

Address for correspondence and reprints: William Navidi, Department of Mathematics, University of Southern California, Los Angeles, California 90089-1113.

© 1991 by The University of Chicago. All rights reserved.
0737-4038/91/0801-0008$02.00
In the present paper, we will discuss two methods which use linear invariants. One is the method described by Cavender (1989), which relies on the asymptotic normality of the data. We will refer to this procedure as the asymptotic method of linear invariants. This method will be shown to contain Lake's (1987) method of evolutionary parsimony as a special case. We will show that Cavender's asymptotic method of linear invariants is essentially a maximum-likelihood technique. Therefore this method can be viewed as a way to base inferences on the likelihood function without assuming identical evolutionary processes in the various sites. We will also discuss an exact binomial version of evolutionary parsimony, similar to that found in the work of Holmquist et al. (1988). We will describe a one-sided test which is appropriate under very general assumptions.

The organization of the present paper is as follows: After describing the general model, we compare the features of the point-estimate type of data analysis with those of the confidence-set approach. Then we discuss in turn the likelihood-ratio method, the evolutionary-parsimony method, and the asymptotic method of linear invariants, describing in detail both point-estimate and confidence-set procedures. We then give a numerical example. Finally we summarize our results, comparing the asymptotic method of linear invariants to the likelihood-ratio method.

The General Model

We present the general model in the context of four-species unrooted trees (see fig. 1). It is easily extended to larger trees and to rooted trees. This model is also described by Navidi et al. (submitted). The species are labeled I–IV. Figure 1 shows one of the three possible trees, which we will refer to throughout as tree 1. Tree 2 is obtained from figure 1 by interchanging species II and III, and tree 3 is obtained by interchanging species II and IV.

We assume we have on hand four aligned nucleic acid sequences. Restrict attention to a single site on the molecule. Since the tree is unrooted, the root could be in any branch. Thus, as we travel along a branch, we cannot determine whether we are moving forward or backward in time. We therefore arbitrarily choose one of the interior nodes of the tree (a node labeled with an asterisk) as the initial node for the tree and specify a direction (indicated by arrows) for each branch (see figure 1). We ignore any effects due to insertions and deletions and assume that all variation in the sequences arises from transitions and transversions. We also ignore any effects due to the alignment process. We make the following assumptions:

![Figure 1: Four-species unrooted tree](image-url)
Assumption 1. The evolutionary processes in the various sites act independently of one another.

Assumption 2. All the sites evolve according to the same tree.

Assumption 3. Given the base at a site at an internal node, the two collections of bases formed by removing that site are distributed independently of each other.

Assumption 3 is a weaker form of the Markov hypothesis, which says that, given the present state of a site, the future evolution is independent of the past. It is similar to assumption 2 of Barry and Hartigan (1987, p. 200) and to assumption 2 of Felsenstein (1983, p. 248).

Let \( r = \{r(A), r(G), r(C), r(U)\} \), where \( r(A), r(G), r(C), \) and \( r(U) \) are the probabilities of observing \( A, G, C, \) and \( U \), respectively, at the initial node. For each branch \( i \), define the substitution probability \( m_i(A,G) \) to be the conditional probability of observing a \( G \) at the ending node of the branch, given that an \( A \) was present at the beginning node. For each of the other 15 pairs of bases \( A, G, C, \) and \( U \), make a similar definition. The Markov matrix \( M_i \) for the \( i \)th branch is given by

\[
M_i = \begin{bmatrix}
m_i(A,A) & m_i(A,G) & m_i(A,C) & m_i(A,U) \\
m_i(G,A) & m_i(G,G) & m_i(G,C) & m_i(G,U) \\
m_i(C,A) & m_i(C,G) & m_i(C,C) & m_i(C,U) \\
m_i(U,A) & m_i(U,G) & m_i(U,C) & m_i(U,U)
\end{bmatrix}
\]

The elements of \( M_i \) are numbers between 0 and 1, and the row sums are equal to 1. This means that 12 parameters must be specified before the matrix is known. Since the components of \( r \) also sum to 1, a total of 63 parameters (5 branches \( \times \) 12 parameters/branch + 3 parameters at the initial node) are needed to determine the model. Additional constraints may be imposed on the parameters if desired. This will be necessary for linear invariants to exist. We require that any constraints be the same at all sites.

Denote the base at the beginning node of branch 5 by \( b_5 \) and denote the base at the end node of branch 5 by \( b_6 \). Then the probability of observing bases \( b_1, b_2, b_3, \) and \( b_4 \) at nodes 1, 2, 3, and 4, respectively, is

\[
\sum_{b_5} \sum_{b_6} r(b_5) m_5(b_5,b_6) m_3(b_6,b_3) m_4(b_6,b_4) m_1(b_2,b_1) m_2(b_5,b_2)
\]

This model describes the evolutionary process at a single site. If the values of the parameters are the same in all sites, we say that the sites are identically distributed.

Let \( N \) denote the number of sites. At each site, each species I-IV contributes an \( A, G, C, \) or \( U \). Thus, each site yields an ordered quartet of bases which falls into one of 256 categories: \( \text{AAAA}, \ldots, \text{UUUU} \). The position in the quartet depends on the species; for example, the quartet \( \text{AGCU} \) means that species I-IV have bases \( A, G, C, \) and \( U \), respectively, at that site. The probability that a given site will be occupied by a given quartet is determined both by a topology of the tree and by the values of the parameters in \( r \) and the \( M_i \). Number the 256 quartets from 1 to 256. Denote by \( p_k \) the 256-dimensional column vector whose \( j \)th component, \( p_{kj} \), is the probability that site \( k \) will be occupied by quartet \( j \). Equation (2) shows how to calculate \( p_k \) in terms of \( r \) and the \( M_i \). Denote by \( \bar{p} \) the average \( \frac{1}{N} \sum_{k=1}^{N} p_k \). Each component of \( \bar{p} \) is in some
sense the "average" probability that a site will be occupied by a certain quartet. If the sites are identically distributed, then \( p_k \) will be the same for each site \( k \), and \( \tilde{p} \) will also take on this common value. Denote by \( X \) the 256-dimensional vector whose \( j \)th component is the number of sites occupied by quartet \( j \). Then \( E(X) = N\tilde{p} \).

Restrict attention to a particular tree \( \tau \) and to a particular site \( k \). Equation (2) generates all possible values of the components of \( p_k \) as the elements of \( r \) and the \( M_i \) run through their allowable values. Denote by \( P^*_\tau \) the set of all possible values of \( p_k \).

Since any constraints on the parameters are the same at all sites, the set \( P^*_\tau \) is the same at each site. Denote by \( P_\tau \) the set of all possible values of \( \tilde{p} = E(X)/N \). If the sites are identically distributed, \( p_k = \tilde{p} \), so \( P_\tau = P^*_\tau \). Otherwise \( P_\tau \) is the set of all averages on \( N \) vectors chosen from \( P^*_\tau \), so \( P_\tau \) is a larger set.

Substitution Probabilities versus Substitution Rates

In the model described above, the parameters represent substitution probabilities. Many important aspects of molecular evolution are more appropriately studied with models involving substitution rates. For example, models which are designed to estimate the expected number of substitutions occurring at a site during a given time are more naturally based on substitution rates. Substitution probabilities are less desirable for this task because they do not directly give information about the number of substitutions that may have created an observed difference. Models involving substitution rates and elapsed times have been used to estimate times between speciation events. Substitution probabilities do not offer as direct an estimation procedure.

A matrix \( Q \) of substitution rates, together with a length of time \( t \), determine a matrix of substitution probabilities by \( e^{tQ} \). In contrast, there exist matrices of substitution probabilities which are not generated by any rate matrix. Thus models involving substitution probabilities are somewhat more general.

Models based on substitution probabilities seem to be sufficient for the task of estimating phylogenies from aligned nucleic acid sequences. The probability distribution of the vector \( X \) is determined by the substitution probabilities, through equation (2). Therefore statistical methods which use \( X \) as the data, such as maximum-likelihood methods and invariant methods, can be based on models whose parameters represent substitution probabilities. Models based on substitution rates and on elapsed times can also work—but only because they determine the substitution probabilities. It seems easier to work directly with the substitution probabilities themselves.

Point Estimates and Confidence Sets

Data analysis for the purpose of estimating phylogenies may pursue one of two approaches. One approach is to find the single tree which best fits the data, to serve as the estimate of the correct phylogeny. An analysis of this type may be called a point estimate approach. Such an analysis may be modified by the inclusion of a rule that, if either no tree or more than one tree fits the data well, then no estimate will be made. A disadvantage of the point-estimate approach is that its error rate depends on the true values of the parameters. Invariably, there will be some values of the parameters for which the procedure is very accurate and others for which it is not. For this reason it is often difficult to determine the accuracy of the procedure in a specific instance.

An alternative approach is to classify each tree in a collection as being plausible or implausible, rather than to single out a best tree. A procedure of this type may be described as a confidence-set approach. In this approach, for each tree \( \tau \), a statistical test is made of the hypothesis that \( \tau \) is the correct tree. Those trees not rejected at level \( \alpha \) are deemed plausible. If the initial collection of trees is known to contain the
correct one, the set of plausible trees is a $1 - \alpha$ confidence set for the correct tree. In contrast to the point-estimate approach, the error rate for the confidence-set procedure is always known. The proportion of confidence sets which fail to contain the correct tree is $\alpha$.

In practice, point-estimate procedures and confidence-set procedures require similar calculations, so it is reasonable to do both. Both likelihood techniques and linear-invariant methods, including evolutionary parsimony, can be used to perform analyses of either type.

**The Likelihood-Ratio Method**

The use of maximum-likelihood techniques in phylogenetic inference is now common. For example, Felsenstein (1981) proposed a model in which the evolutionary process in each branch of the tree is governed by a single rate parameter. Other likelihood methods have been proposed since; among the most flexible is Barry and Hartigan's (1987) "maximum average likelihood" approach. These methods provide techniques to compute the maximum of the likelihood function for a given tree. The tree whose maximum likelihood is the largest is a point estimate of the correct tree. It has not been clear how to use the likelihood function to find a confidence set. This requires a hypothesis test to determine whether a given tree is compatible with the data. The likelihood-ratio method, described below, yields such a test. It provides a measure of the statistical significance of the maximum value of the likelihood function for any tree. Thus, given the maximum likelihood for each of several competing trees, each tree can be tested in turn and classified as being either plausible or implausible. When the number of species is small enough so that every tree can be tested, the plausible trees form a $1 - \alpha$ confidence set.

We now describe the likelihood-ratio method in terms of the general model given above. A description from a somewhat more mathematical point of view is given by Navidi et al. (submitted). Although the discussion is given in the context of four-species trees, the method can in principle be extended to trees of any size. In addition to assumptions 1-3, we impose two additional assumptions on the model.

**Assumption 4.** The sites are identically distributed.

**Assumption 5.** The values of the parameters in $\mathbf{r}$ and in the $\mathbf{M}_i$ are equal neither to 0 nor to 1.

Assumption 5 states that the true values of the parameters lie in the interior of the parameter space. It is necessary in order to ensure that the likelihood-ratio statistic has the correct asymptotic distribution. It is harmless to make this assumption, since in practice the behavior of a model when some parameters lie on the boundary is indistinguishable from its behavior when those same parameters are very near but not on the boundary.

Fix a tree, $\tau$, to be tested. As before, let $\mathbf{X}$ refer to the vector of counts of the 256 quartets, let $\bar{\mathbf{p}} = E(\mathbf{X})/N$ be the vector of expected frequencies, and denote the $i$th coordinate of $\bar{\mathbf{p}}$ by $\bar{p}_i$. Recall that $P_\tau$ is the set of all possible values of $\bar{\mathbf{p}}$ when $\tau$ is the correct tree. We perform a test of the null hypothesis

$$H_0: \bar{\mathbf{p}} \in P_\tau.$$  (3)

Since the sites are identically distributed, $\mathbf{X}$ has the multinomial distribution with
parameters $N$ and $p$. The likelihood function is the density of $X$ considered as a function of the parameter $p$.

$$l(\mathbf{p}) = \frac{N!}{X!} \prod_{i=1}^{256} \mathbf{p}_i^{X_i}. \quad (4)$$

More useful is the log likelihood, the logarithm of $l$ when the constant $N! / \prod_{i=1}^{N} X_i!$ is ignored:

$$L(\mathbf{p}) = \sum_{i=1}^{256} X_i \log \mathbf{p}_i. \quad (5)$$

The maximum-likelihood estimator (MLE) of $\mathbf{p}$ is the vector $\mathbf{\hat{p}} \in P_r$, maximizing $L$. We assume such a vector exists. Methods for computing the MLE are based on the fact that all the vectors in $P_r^+ (= P_r)$ are generated by equation (2) as the parameters in $r$ and in the $M_i$ run through their allowable values. See Barry and Hartigan (1987) for an algorithm to compute $\mathbf{\hat{p}}$ for the full model in which each matrix has 12 free parameters. Felsenstein (1981) describes an algorithm for the case in which each matrix is determined by a single rate parameter. The value $L(\mathbf{\hat{p}})$ is called the constrained maximum likelihood. It is the maximum value taken by the log likelihood by using values of the parameters allowable under the constraints of the model.

Previous efforts to use maximum likelihood to test the hypothesis that a tree $\tau$ is correct have involved expressing $\tau$ as a subcase of a more general tree $\tau'$, where $\tau'$ is known to be correct. The tree $\tau$ can be obtained from $\tau'$ by placing some constraints on parameters, e.g., setting their values equal to zero. See the work of Felsenstein (1988) for a discussion of this idea. The principle involved is that, if the subtree $\tau$ is in fact correct, then the statistic $2[L(\mathbf{\hat{p}}_{\tau'}) - L(\mathbf{\hat{p}}_{\tau})]$ has, asymptotically, a $\chi^2$ distribution with degrees of freedom equal to the number of parameter constraints.

This principle is more generally applicable. It is not necessary to express $\tau$ as a subcase of a correct tree. We need only express $\tau$ as a subcase of a model for which it is known that the correct tree is also a subcase. The most general model, of which every tree is a subcase, is the model putting no constraints at all on $\mathbf{p}$, except that its components sum to 1. This is known as the unconstrained or saturated model.

With no constraints on $\mathbf{p}$, the MLE of $\mathbf{p}$ is $X/N$. The unconstrained maximum likelihood is $L(X/N) = \sum_{i=1}^{256} X_i \log X_i/N$, where $X_i \log X_i/N$ is taken to be 0 if $X_i = 0$. The quantity

$$\lambda_r = 2[L(X/N) - L(\mathbf{\hat{p}}_{\tau})] \quad (6)$$

is the generalized likelihood-ratio statistic for testing $H_0: \mathbf{\bar{p}} \in P_r$.

Define $d$ as 255 – the number of free parameters in the model. Under assumptions 1–5, if $\mathbf{\hat{p}}_{\tau}$ is a local maximum of $L$, then the asymptotic null distribution of $\lambda_r$ is $\chi^2_d$ (see Rao 1973, pp. 418–419).

When more than five or six species are to be considered simultaneously, there are too many possible trees to test each one. In such cases, the likelihood-ratio statistic can be used in conjunction with various search techniques to find plausible trees for larger numbers of species. Several useful such techniques have been proposed. For example, Felsenstein (1981) describes a search strategy which begins with a two-species
tree, then adds species one by one, placing new species on the branches which yield the greatest likelihood. His discussion is given in the context of a model in which the matrix for each branch of the tree is determined by a single rate parameter. Barry and Hartigan (1987) describe an application of this technique to the fully parameterized model. Given a tree or set of trees which have been found by these or other search procedures, the likelihood-ratio statistic can be computed to test the plausibility of each one.

We now discuss the behavior of the procedure when the assumption of identically distributed sites is dropped. Recall that $p_k$ is the vector of probabilities associated with the $k$th site and that the collection of possible values of $p_k$, denoted $P^*_k$, is the same for each site $k$. The vectors $p_k \in P^*_k$ can be expressed in terms of the $\leq 63$ free parameters in $\mathbf{r}$ and the $\mathbf{M}_i$, through equation (2). When the sites are not identically distributed, $P_\tau$ is a larger set than $P^*_\tau$, and the vectors $\mathbf{p} \in P_\tau$ are not expressible in this way. We can still compute $\lambda_\tau$ with equation (6), using the quantity $\mathbf{p} \in P^*_\tau$ which maximizes the likelihood. Since we will then be maximizing over too small a set, $L(\mathbf{p})$ will tend to be less than the maximum of $L$ taken over $P_\tau$, so $\lambda_\tau$ will tend to be larger under the null hypothesis than under the $\chi^2$ distribution given above. Thus we will be more likely to reject the null hypothesis than the presumed level of the test would indicate.

**Evolutionary Parsimony**

The method of evolutionary parsimony, developed by Lake (1987), is a procedure applicable to four-species unrooted trees (see fig. 1). Corresponding to each tree is a pair of 256-dimensional column vectors called *linear invariants*. Each tree shares an invariant with each other, so there are three linear invariants in all. The correct tree has the property that both its invariants are orthogonal to $\mathbf{p}$.

Lake (1987) used the term "invariant" to refer to the test statistics generated by his method. Cavender (1989) used the term in a slightly different way. The following definition, which we will use, is equivalent to Cavender's:

**Definition 1.** A vector is a linear invariant for the tree $\tau$ if it is orthogonal to every vector in $P_\tau$.

For Lake's method to work, we need assumptions 1-3. We also need some parameter constraints. No restrictions need be put on $\mathbf{r}$ or on the matrix for the middle branch; whereas the following assumption is needed for the outer branches:

**Assumption 6.** The matrices for the outer branches are of the following form:

$$
\begin{align*}
A & \quad \begin{bmatrix} a & b & c & d \\ e & f & g & g \\ h & i & j & j \\ k & k & l & m \\ n & n & p & q \end{bmatrix},
\end{align*}
$$

This means that when a transversion occurs in an outer branch it is equally likely to be either of the two possible ones. There are no other constraints. We do not need assumption 4 or assumption 5. The method is valid regardless of whether the sites are identically distributed.

Jin and Nei (1990) describe for the model a parameterization in which the Markov matrices satisfy assumption 6 yet Lake's method fails. Their matrices give substitution probabilities over a small fraction (e.g., $1/50$) of a branch and are approximations of
Methods of Phylogenetic Inference 135

substitution-rate matrices. In this parameterization, additional conditions are needed, so that the collection of allowable matrices is closed under multiplication. As pointed out by Jin and Nei, sufficient conditions are given by Cavender (1989). These conditions are mentioned below. Assumption 6 is sufficient, however, when the matrices give substitution probabilities over the whole branch, as we have done here.

We now describe the three vectors which are the linear invariants: let $v_1$ be the 256-dimensional vector with a 1 in the positions corresponding to the quartets AACC, AAUU, GGCC, GGUU, CCAA, CCGG, UUAA, UUGG, AGCU, AGUC, GACU, GAUC, CUAG, CUGA, UCAA, UCAG, AAUC, GGCU, GGUC, CCAG, CGCA, UUAG, and UUGA; and 0 elsewhere.

Let $v_2$ be the 256-dimensional vector with a 1 in the positions corresponding to the quartets ACAC, AUAU, GCGC, GUGU, CACA, CGCG, UAUA, UGUG, ACGU, AUGC, GACA, CAUA, CGUG, UACG, UGCG, AUCG, AUAC, GCGG, GUGC, CAGC, CGAC, CAGA, UAGU, and UGUA; and 0 elsewhere.

Let $v_3$ be the 256-dimensional vector with a 1 in the positions corresponding to the quartets ACCA, AUUA, GCCG, GUUG, CAAC, CGGC, UAAU, UGGU, ACUG, AUGC, GCAC, CAGU, CGAU, UAGC, and UGAC; a -1 in the positions corresponding to the quartets ACCG, AUUG, GCCA, GUUA, CAAA, CGGU, UAGU, and UGUA; and 0 elsewhere.

Denote by $v_i^T$ the transpose of $v_i$, i.e., $v_i$ considered as a row vector. Then $v_i$ is orthogonal to $\overline{p}$ if $v_i^T \overline{p} = 0$.

Lake (1987) showed that, if tree 1 is correct, then $v_1^T\overline{p} = 0$ and $v_2^T\overline{p} = 0$. Thus $v_2$ and $v_3$ are linear invariants for tree 1. Similarly, $v_1$ and $v_3$ are invariants for tree 2, and $v_1$ and $v_2$ are invariants for tree 3. For $i = 1, 2, 3$, let $P_i$ be the sum of the components of $X$ in those positions where $v_i = 1$, and let $Q_i$ be the sum of the components of $X$ in those positions where $v_i = -1$, so $v_i^T X = P_i - Q_i$.

We now describe three point-estimate procedures and a confidence-set procedure using the setup above. The first point-estimate approach is due to Lake (1987). For $i = 1, 2, 3$, we test the hypothesis

$$H_0: v_i^T \overline{p} = 0.$$  \hspace{1cm} (8)

The hypothesis $H_0$ is true when tree $i$ is incorrect and, when highly exceptional cases are disregarded, is false when tree $i$ is correct.

The test statistic proposed by Lake is

$$S_i = \frac{(P_i - Q_i)^2}{P_i + Q_i}.$$  \hspace{1cm} (9)

This statistic can be justified as follows: The quantity $P_i - Q_i = v_i^T X$, so $E(P_i - Q_i) = E(v_i^T X) = N\overline{v_i^T \overline{p}} = 0$ under $H_0$. The variance of $P_i - Q_i$ is $N\overline{v_i^T |\overline{p}|}$, where $|v_i^T|$ denotes the vector whose components are the absolute values of the components of $v_i^T$. We estimate $N\overline{|v_i^T| \overline{p}}$ by substituting $X$ for $N\overline{p}$, obtaining $|v_i^T| X = P_i + Q_i$. When $N$ is large, $P_i - Q_i$ is approximately normally distributed, so, under $H_0$, $S_i$ has an approximate $\chi^2_1$ distribution. Lake suggested computing the three quantities $S_1, S_2,$
and $S_3$ and declaring a tree $i$ to be correct if $S_i$ was significant at the 5% level and if $S_j$, for $j \neq i$, was not. Otherwise no tree would be preferred.

Holmquist et al. (1988) pointed out that, if tree $i$ is incorrect, then conditional on $P_i + Q_i$, $P_i$ has the binomial distribution with number of trials equal to $P_i + Q_i$ and success probability $\pi = \frac{1}{2}$. Therefore an exact binomial test can be made, avoiding the $\chi^2$ approximation. Of course, a test which has conditional level $\alpha$ for each value of $P_i + Q_i$ is an unconditional level-$\alpha$ test as well.

We now make an assumption which will ensure that $\pi > \frac{1}{2}$ when tree $i$ is correct. Then a one-sided test will be appropriate, which will increase power considerably.

Assumption 7. In each branch of the tree, the probability of no difference between the nodes at either end is greater than the probability of a transition difference.

Theorem 1. Under assumptions 1–3, 6, and 7, if tree $i$ is correct, then conditional on $P_i + Q_i$, $P_i$ has the binomial distribution with number of trials equal to $P_i + Q_i$ and success probability $\pi > \frac{1}{2}$. (The proof is in the Appendix.)

We now define our second point-estimate procedure for choosing a tree using Lake's invariants.

Procedure A:

A1. Choose a critical value $\alpha$ and compute $P_1$, $P_2$, and $P_3$.

A2. Compute the one-sided upper-tail significance level of $P_i$ by using the binomial distribution with parameters $P_i + Q_i$ and $\frac{1}{2}$.

A3. Declare a tree to be correct if its significance level is $< \alpha$ while the significance levels of the other two trees are $> \alpha$. Make no decision otherwise.

This is a one-sided exact version of Lake's (1987) procedure, which we have described above. Alternatively, we could simply choose the tree with the lowest significance level.

It has been suggested (Li 1989) that when Lake's method is used with significance level $\alpha$ the true significance level is $3\alpha$, since three tests are being made. The significance level measures the frequency with which a true null hypothesis is rejected. In the context of tree-selection procedures, such as evolutionary parsimony, another error rate seems to be more important. This is the frequency with which the procedure chooses an incorrect tree. This frequency depends on the true values of the parameters and is difficult to determine. In general, it need not be well estimated by $3\alpha$.

A procedure which yields confidence sets for the correct tree can be constructed by considering sums of pairs of the $P_i$. Define

$$Y_1 = P_2 + P_3; Y_2 = P_1 + P_3; Y_3 = P_1 + P_2 \tag{10}$$

and

$$Z_1 = Q_2 + Q_3; Z_2 = Q_1 + Q_3; Z_3 = Q_1 + Q_2. \tag{11}$$

If tree $i$ is correct, then conditional on $Y_i + Z_i$, $Y_i$ has the binomial distribution with parameters $Y_i + Z_i$ and $\frac{1}{2}$. If tree $i$ is incorrect, then $Y_i$ is conditionally binomial with success probability $> \frac{1}{2}$. For each tree $i$ we may compute the one-sided upper-tail significance level of $Y_i$ by using the binomial distribution with parameters $Y_i + Z_i$ and
The collection of trees for which this significance level is <\(\alpha\) is an exact level \(1 - \alpha\) confidence set for the correct tree.

A reviewer has pointed out that, if we define \(Y_i' = P_2 + Q_3; Y_i = P_1 + Q_3; Y_i' = P_1 + Q_2\) and \(Z_i' = Q_2 + P_3; Z_i = Q_1 + P_3; Z_i = Q_1 + P_2\), then, if tree \(i\) is correct, it is the case that, conditional on \(Y_i' + Z_i'\) and \(\frac{Y_i}{Z_i}\), \(Y_i\) has the binomial distribution with parameters \(Y_i' + Z_i'\) and \(\frac{Y_i}{Z_i}\). If tree \(i\) is incorrect, then \(Y_i\) is conditionally binomial with success probability \(\neq \frac{Y_i}{Z_i}\). For each tree \(i\) we may compute the two-sided significance level of \(Y_i\) by using the binomial distribution with parameters \(Y_i' + Z_i'\) and \(\frac{Y_i}{Z_i}\).

We may combine the procedure above with the procedure based on equations (10) and (11) by performing each and then rejecting tree \(i\) if either procedure rejects it at level \(\alpha/2\). Some preliminary calculations indicate that, in some situations, combining the two procedures may increase power by \(\sim 10\%\) over that of the procedure based on equations (10) and (11) alone. In other situations, the power of the combined procedure seems to be about the same.

The statistics \(Y_i, Z_i\) can be used to formulate another point-estimate procedure for choosing the correct tree, as described below:

**Procedure B:**

1. **B1.** Choose a critical value \(\alpha\) and compute \(Y_1, Y_2,\) and \(Y_3\).
2. **B2.** Compute the one-sided upper-tail significance level of \(Y_i\) by using the binomial distribution with parameters \(Y_i + Z_i\) and \(\frac{Y_i}{Z_i}\).
3. **B3.** Declare tree \(i\) to be correct if \(Y_i\) is not significant at level \(\alpha\) whereas \(Y_j\) for \(j \neq i\) is significant. Make no decision otherwise.

Again, an obvious modification is to choose the tree with the largest significance level.

This procedure can also be combined with the two-sided procedure based on \(Y_i'\) and \(Z_i'\). Preliminary simulation results suggest that procedures A and B are about equally effective in selecting the correct tree.

We now describe another confidence-set procedure, valid when the number of sites is large, which can be extended to a more general setting. Fix in mind a tree \(i\) and let \(V_i\) be the \(256 \times 2\) matrix whose columns are the invariants for the tree \(i\). Thus \(V_i\) is the matrix whose columns are \(v_j\) and \(v_k\), where \(i \neq j\) and \(i \neq k\). Denote by \(V_i^T\) the transpose of \(V_i\), the \(2 \times 256\) matrix whose rows are the columns of \(V_i\). Then, if tree \(i\) is correct, \(V_i^T \tilde{\rho} = 0\). We seek a test statistic for the hypothesis

\[
H_0: V_i^T \tilde{\rho} = 0. \tag{12}
\]

The quantity \(V_i^T X\) is the two-dimensional column vector whose components are \(P_j - Q_i\) and \(P_k - Q_k\). It is distributed approximately bivariate normal, with mean \(E(V_i^T X) = N V_i^T \tilde{\rho} = 0\) under \(H_0\). The covariance matrix of \(V_i^T X\) is \(V_i^T \Sigma V_i\), where \(\Sigma\) is the \(256 \times 256\)-dimensional covariance matrix of \(X\). We must estimate \(V_i^T \Sigma V_i\). In general this is done by first estimating \(\Sigma\), but in this case we can do it more directly. The quantities \(P_j - Q_i\) and \(P_k - Q_k\) have variances \(N|v_j^T| \tilde{\rho}\) and \(N|v_k^T| \tilde{\rho}\), respectively, and we show in the appendix that they are uncorrelated. Therefore the matrix \(V_i^T \Sigma V_i\) is the \(2 \times 2\)-dimensional matrix whose main diagonal elements are \(N|v_j^T| \tilde{\rho}\) and \(N|v_k^T| \tilde{\rho}\) and whose off-diagonal elements are 0. We estimate \(N|v_j^T| \tilde{\rho}\) with \(P_j - Q_j\) and \(N|v_k^T| \tilde{\rho}\) with \(P_k + Q_k\). Substituting, we obtain the estimated covariance matrix, which we denote \(V_i^T \Sigma V_i\). To compute the test statistic, we must compute the inverse of this matrix. This is easily done. The matrix \((V_i^T \Sigma V_i)^{-1}\) has elements \(1/(P_j + Q_j)\) and \(1/(P_k + Q_k)\) on the main diagonal and 0 on the off diagonal. Define the test statistic
Under \( H_0 \), \( T_i \) has an approximate \( \chi^2 \) distribution. In fact, \( T_i = S_j + S_k \), the sum of two of Lake's statistics. See equation (9). The level \( 1 - \alpha \) confidence set is determined by including those trees \( i \) for which \( T_i \) is not significant at level \( \alpha \).

This confidence procedure is itself not very useful, since the exact procedure given above is more accurate and easier to compute. What is of value is its method of construction, which will be used below to develop a very general procedure.

### The Asymptotic Method of Linear Invariants

When appropriate constraints are placed on the matrix parameters, it is in principle possible to compute invariants for trees of any size. Cavender (1989) describes a method for calculating all linear invariants for four-species-rooted trees with six branches. In principle, his method can be generalized to larger trees, although the amount of computation becomes too great for trees with more than about five species. For linear invariants to exist, some constraints must be put on the parameters. For example, Cavender (1989) considers matrices of the form of matrix (7) with two additional constraints:

\[
e + h = f + i, \text{ and } l + p = m = q. \tag{14}
\]

The set of all matrices of this form is closed under matrix multiplication (and is thus a semigroup). No strictly larger semigroup than this is useful, since then there are no linear invariants (Cavender, submitted). Models in which the probabilities of transversions are not balanced are also feasible (Cavender 1989).

Fix in mind a tree \( \tau \) to be tested. We presume that appropriate parameter constraints are in place and that a number of linear invariants for \( \tau \) have been calculated. Let \( m \) represent this number. It is not necessary to calculate all of the linear invariants; any subset will do. However, each invariant adds 1 degree of freedom to the distribution of the test statistic, so more invariants should generally result in a more powerful test.

Let \( V_i \) be the \( 256 \times m \) matrix whose columns are linear invariants for \( \tau \). We construct a test statistic for testing the null hypothesis

\[
\mathcal{H}_0: V_i^T \tilde{\rho} = 0. \tag{15}
\]

The hypothesis \( \mathcal{H}_0 \) is true if \( \tau \) is the correct tree. The test statistic will be of the form of equation (13). Its construction is analogous to the one carried out at the end of the previous section. When the number of sites is large, the quantity \( V_i^T X \) is distributed approximately \( m \)-variate normal, with mean \( \mathbb{E}(V_i^T X) = N V_i^T \tilde{\rho} = 0 \) under \( \mathcal{H}_0 \). As before, let \( \Sigma \) denote the covariance matrix of \( X \). We describe a procedure to estimate \( V_i^T \Sigma V_i \), the covariance matrix of \( V_i^T X \). In this case, we will estimate \( V_i^T \Sigma V_i \) by estimating \( \hat{\Sigma} \). Consider the case in which the sites are identically distributed, although the procedure works equally well when they are not. Denote the \( ij \)th element of \( \Sigma \) by \( \sigma_{ij} \). Then, if the sites are identically distributed,

\[
\sigma_{ij} = -N \tilde{p}_i \tilde{p}_j \text{ if } i \neq j, \text{ and } \sigma_{ii} = N \tilde{p}_i (1 - \tilde{p}_i). \tag{16}
\]

The estimator \( \hat{\Sigma} \) will be based on an estimator \( \hat{\rho} \) of \( \tilde{\rho} \). Any consistent estimator of \( \hat{\rho} \)
will work asymptotically. One possible choice is \( \hat{\rho} = X/N \), but we will use \( \hat{\rho} = X/N - V_i(V_i^T V_i)^{-1} V_i^T X/N \), subtracting from \( X/N \) its projection onto the space spanned by the columns of \( V_i \). This should be more accurate, and it seems to produce less complicated expressions for the test statistic. This procedure will give Lake's test statistics (9) under the assumptions of evolutionary parsimony. Denote the \( ij \)th element of \( \Sigma \) by \( \hat{\sigma}_{ij} \). Now define \( \Sigma \) as follows:

\[
\hat{\sigma}_{ij} = -N\hat{p}_i\hat{p}_j \text{ if } i \neq j, \text{ and } \hat{\sigma}_{ii} = N\hat{p}_i(1 - \hat{p}_i). \tag{17}
\]

The expression on the right hand side of equation (13) now yields a test statistic for the null hypothesis (15). We denote this test statistic by \( T \). It has been shown by Navidi et al. (submitted) that under assumptions 1–3 the asymptotic null distribution of \( T \) is \( \chi^2_n \), regardless of whether the sites are identically distributed. Now, given a collection of trees, we can compute \( T \) for each one. The trees for which \( T \) is not significant at level \( \alpha \) are plausible candidates for the correct tree.

In practice, with samples of moderate size, some problems may be encountered with the above procedure. Cavender (1989) reports getting negative components in \( \hat{\rho} \) with a sample size of 1,095 sites. He replaced the negative values with zeroes, which is probably the best one can do under the circumstances. This indicates that unless the number of sites is quite large the asymptotic approximation should not be trusted, and so the null distribution should be estimated with a simulation.

It turns out that there is a great similarity between the asymptotic method of linear invariants, which is described above, and the likelihood-ratio method. To see this, again fix in mind a tree \( \tau \), and let \( V_\tau \) be a matrix whose columns are linear invariants. Let \( V_\tau \) be the collection of all vectors \( w \) for which \( V_\tau w = 0 \). The null hypothesis (15) can now be written as

\[
H_0: \hat{\rho} \in V_\tau. \tag{18}
\]

It follows from definition 1 that \( P_\tau \subseteq V_\tau \). Comparing hypothesis (18) with hypothesis (3) shows that the asymptotic method of linear invariants differs from the likelihood-ratio method in that it tests whether \( \hat{\rho} \) belongs to the larger set \( V_\tau \) rather than to \( P_\tau \).

The next result states that this is essentially the only difference between the two methods.

**Theorem 2.** Let \( \tau \) denote the correct tree. Let \( \hat{\rho}_V \) be the vector in \( V_\tau \) maximizing \( L \). Let \( T \) be the test statistic for the asymptotic method of linear invariants. Then, under assumptions 1–3 and 5, as the number of sites grows large, the difference \( T - 2[L(X/N) - L(\hat{\rho}_V)] \to 0. \)

See Navidi et al. (submitted) for a proof. Theorem 2 shows that this method of linear invariants is asymptotically equivalent to the likelihood-ratio method in which the likelihood is maximized over \( V_\tau \) instead of over \( P_\tau \). Thus the asymptotic method of linear invariants is essentially a maximum-likelihood technique.

**A Numerical Example**

To provide an example of the methods discussed above, we extracted four aligned sequences from the collection of small-subunit RNAs (Dams et al. 1988). The sequences are I, *Sulfolobus solfataricus* (archaebacteria); II, *Halobacterium salinarium* (archaebacteria); III, *Escherichia coli* (eubacteria); and IV, *Homo sapiens* (eukaryota). This set of four aligned sequences contains 1,352 positions without insertions or dele-
tions. We analyze the data first with evolutionary parsimony, then with the likelihood-ratio method.

To perform an evolutionary parsimony analysis, the following values were calculated:

\[ P_1 = 38, \quad P_2 = 21, \quad P_3 = 25, \quad Q_1 = 30, \quad Q_2 = 12, \quad Q_3 = 15 \]

and

\[ Y_1 = 46, \quad Y_2 = 63, \quad Y_3 = 59, \quad Z_1 = 27, \quad Z_2 = 45, \quad Z_3 = 42 \]

We compute a \( p \) value for each tree. First, under evolutionary parsimony procedure A, the \( p \) value for tree 1 is found by computing the probability that a quantity distributed binomially with number of trials equal to \( P_1 + Q_1 = 68 \) and with success probability \( \frac{1}{2} \) is greater than or equal to \( P_1 = 38 \). This \( p \) value is 0.1981. Similar calculations for trees 2 and 3 yield \( p \) values of 0.0814 and 0.0769, respectively. With these data, the sample size is large enough so that the normal approximation to the binomial would have been appropriate, although we give the exact \( p \) values. In procedure A small \( p \) values are characteristic of the correct tree. Thus evolutionary-parsimony procedure A slightly favors tree 3, which clusters I with IV and II with III. This is the tree proposed by Lake (1988). No tree is significant at the 5% level. The results for procedure B are similar. Here the \( p \) value for tree 1 is found by computing the probability that a quantity distributed binomially with number of trials equal to \( Y_1 + Z_1 = 73 \) and with success probability \( \frac{1}{2} \) is greater than or equal to \( Y_1 = 46 \). This \( p \) value is 0.0172. Similar calculations for trees 2 and 3 yield \( p \) values of 0.0507 and 0.0555, respectively. In this procedure small \( p \) values are characteristic of incorrect trees, so tree 3 is slightly favored. Tree 1 is rejected at the 5% level, and trees 2 and 3 form a 95% confidence set for the correct tree.

To implement the likelihood-ratio method, we use a model of Felsenstein (1981). There are no constraints on \( r \), while the evolutionary process in each branch is described with a single rate parameter. We can describe this model in terms of the general model of section 2 as follows: Denote the rate parameters \( t_1, t_2, t_3, t_4, \) and \( t_5 \). Then the elements of the matrices \( M_i \) are given by

\[ m_i(j,k) = e^{-t_i} \delta_{jk} + (1 - e^{-t_i}) r(k) . \]  

Here \( \delta_{jk} = 1 \) if \( j = k \) and is equal to 0 otherwise. There are a total of eight free parameters, so the asymptotic distribution of the likelihood-ratio statistic has \( 255 - 8 = 247 \) degrees of freedom. The unconstrained maximum likelihood is \( L(X/N) = -5591.057 \). The values of the parameters maximizing the likelihood for tree 1 are

\[ r_1 = 0.229, \quad r_2 = 0.309, \quad r_3 = 0.259, \quad r_4 = 0.203 \]

and

\[ t_1 = 0.211, \quad t_2 = 0.244, \quad t_3 = 0.407, \quad t_4 = 0.652, \quad t_5 = 0.085 . \]

We substitute these quantities into (20) to compute the \( M_i \), then use equation (2) to compute \( \hat{\theta}_i \). The constrained maximum likelihood under tree 1 turns out to be \( L(\hat{\theta}_1) \)
Methods of Phylogenetic Inference

\[ = -5863.775. \] Now the likelihood-ratio statistic (6) for testing tree 1 is \( \lambda_t = 2[(-5591.057 - (-5863.775))] = 545.436. \) The \( p \) value is \( \sim 2.6 \times 10^{-24}. \) The likelihood-ratio method rejects the other two trees just as soundly. Similar calculations for trees 2 and 3 yield likelihood-ratio statistics \( \lambda_t = 550.531 \) for tree 2 and \( \lambda_t = 553.191 \) for tree 3, so the \( p \) values for these trees are slightly smaller than those for tree 1. Tree 1 is thus favored, although every tree is rejected. Tree 1 is the tree proposed by Woese (1987).

The analysis above is intended to illustrate the methods involved and is obviously inadequate for drawing firm conclusions about the correct tree. The small \( p \) values in the evolutionary-parsimony analysis indicate that some deviations from equal transversion probabilities are being picked up in all of the trees, particularly in trees 2 and 3. The very small \( p \) values for the likelihood-ratio test indicate that the model is not describing the data at all well. The problem may be that the parameter constraints are unrealistic, or that the sites are not independent and identically distributed, or both. The likelihood-ratio test is probably quite powerful, in which case it will detect deviations from the model even when they are small.

**Discussion**

We can now compare the asymptotic method of linear invariants with the likelihood-ratio approach. Felsenstein (1988) also provides a very good discussion of some of the ideas mentioned below. The likelihood-ratio approach allows full generality in the model, i.e., 63 free parameters in an unrooted four-species tree, and involves a \( \chi^2 \) test with \( 255 - 63 = 192 \) degrees of freedom. In contrast, linear invariants require parameter constraints. For example, in Cavender's application of the method of linear invariants to a rooted tree with six branches, he allowed 39 free parameters, and the resulting test statistics had either 54 or 68 degrees of freedom, depending on the tree being tested. The likelihood-ratio test in this case would allow 75 free parameters (6 branches \( \times \) 12 parameters/branch + 3 parameters at the initial node) and would have 180 degrees of freedom. In practice, more degrees of freedom usually means more power, so under the assumption of identically distributed sites the likelihood ratio test is likely to be more powerful than the asymptotic method of linear invariants. This leads to more accurate point estimates and smaller confidence sets.

The advantage of linear invariants is the lack of a need for the assumption of identically distributed sites. Whether this is worth the price of extra constraints on the parameters and of a loss of many degrees of freedom depends on how nonidentically the sites are distributed. If in fact the evolutionary process differs greatly from site to site, the likelihood-ratio method may be unreliable, and the asymptotic method of linear invariants may be preferred. On the other hand, if most of the sites evolve somewhat similarly, the extra degrees of freedom in the likelihood-ratio test probably outweigh the falseness of the assumption of identical distributions.

The discussion in the last two paragraphs is applicable when the sample sizes are large enough for the asymptotics to hold. Without further study, it is difficult to give a rule of thumb as to how large such a sample must be. It is likely that the asymptotics do not hold well for small or moderate sample sizes. In practice it is probably best to estimate significance levels by simulations rather than by asymptotic approximations. This does not apply, of course, in the case of evolutionary parsimony, since exact tests are available, the accuracy of which is known for small as well as large sample sizes.
Acknowledgments

This research was supported by the Deutsches Forschungsgemeinschaft, the National Science Foundation, the National Institutes of Health, and the University of Southern California Faculty Research and Innovation Fund. We would like to express our appreciation to Simon Tavaré, whose frequent consultations were invaluable. Thanks also to Walter Fitch, Terry Speed, and Michael Waterman for many helpful suggestions and to an anonymous reviewer for suggesting the statistics \( Y_i \) and \( Z_i \).

APPENDIX

Proof of theorem 1: Direct calculation using equation (2) shows that, under assumption 6, if tree \( i \) is correct, then

\[
\frac{E(P_i - Q_i)}{N} = r(A)m_5(A, C)\left( [m_1(A, A) - m_1(A, G)][m_2(A, A) - m_2(A, G)] \right.
\]
\[
\times [m_3(C, C) - m_3(C, U)][m_4(C, C) - m_4(C, U)]
\]
\[
+ r(A)m_5(A, U)\left( [m_1(A, A) - m_1(A, G)][m_2(A, A) - m_2(A, G)] \right.
\]
\[
\times [m_3(U, C) - m_3(U, U)][m_4(U, C) - m_4(U, U)]
\]
\[
+ r(G)m_5(G, C)\left( [m_1(G, A) - m_1(G, G)][m_2(G, A) - m_2(G, G)] \right.
\]
\[
\times [m_3(C, C) - m_3(C, U)][m_4(C, C) - m_4(C, U)]
\]
\[
+ r(G)m_5(G, U)\left( [m_1(G, A) - m_1(G, G)][m_2(G, A) - m_2(G, G)] \right.
\]
\[
\times [m_3(U, C) - m_3(U, U)][m_4(U, C) - m_4(U, U)]
\]
\[
+ r(C)m_5(C, A)\left( [m_1(C, C) - m_1(C, U)][m_2(C, C) - m_2(C, U)] \right.
\]
\[
\times [m_3(A, A) - m_3(A, G)][m_4(A, A) - m_4(A, G)]
\]
\[
+ r(C)m_5(C, G)\left( [m_1(C, C) - m_1(C, U)][m_2(C, C) - m_2(C, U)] \right.
\]
\[
\times [m_3(G, A) - m_3(G, G)][m_4(G, A) - m_4(G, G)]
\]
\[
+ r(U)m_5(U, A)\left( [m_1(U, C) - m_1(U, U)][m_2(U, C) - m_2(U, U)] \right.
\]
\[
\times [m_3(A, A) - m_3(A, G)][m_4(A, A) - m_4(A, G)]
\]
\[
+ r(U)m_5(U, A)\left( [m_1(U, C) - m_1(U, U)][m_2(U, C) - m_2(U, U)] \right.
\]
\[
\times [m_3(G, A) - m_3(G, G)][m_4(G, A) - m_4(G, G)]
\]
\[
\left. \right)
\]

(A1)

Since \( P_i \) and \( Q_i \) are disjoint sums of multinomial counts, then, conditional on \( P_i + Q_i \), \( P_i \) has the binomial distribution with parameters \( P_i + Q_i \) and \( \pi \), where \( \pi = E(P_i)/E(P_i + Q_i) = 1/2 + E(P_i - Q_i)/2E(P_i + Q_i) \). Equation (22) shows that under assumption 7, when tree \( i \) is correct, \( E(P_i - Q_i) > 0 \), so \( \pi > 1/2 \).

Proof that \( P_j - Q_j \) and \( P_k - Q_k \) are uncorrelated under the null hypothesis (12) is as follows: By definition, \( P_j - Q_j \) and \( P_k - Q_k \) are uncorrelated if \( E[(P_j - Q_j)(P_k - Q_k)] = E[(P_j - Q_j)]E[(P_k - Q_k)] \). Since, under \( H_0 \), \( E[(P_j - Q_j)] = E[(P_k - Q_k)] = 0 \), it suffices to show that \( E[(P_j - Q_j)(P_k - Q_k)] = 0 \).

Focus on the \( i \)th site where \( 1 \leq i \leq N \). Define \( P_{ij} = 1 \) if the quartet at the \( i \)th site corresponds to one of the quartets for which \( v_j = 1 \). Otherwise \( P_{ij} = 0 \). Define \( Q_{{ij}} = 1 \) if the quartet at the \( i \)th site corresponds to one of the quartets for which \( v_j = -1 \).
Otherwise $Q_{ij} = 0$. Define $P_{ik}, Q_{ik}$ similarly. Then at most one of $P_{ij}, Q_{ij}, P_{ik}, Q_{ik}$ is nonzero, so $(P_{ij} - Q_{ij})(P_{ik} - Q_{ik}) = 0$. Now $P_{j} - Q_{j} = \sum_{i=1}^{N} (P_{ij} - Q_{ij})$, and $P_{k} - Q_{k} = \sum_{m=1}^{N} (P_{mk} - Q_{mk})$. Therefore $E[(P_{j} - Q_{j})(P_{k} - Q_{k})] = \sum_{i=1}^{N} \sum_{m=1}^{N} E[(P_{ij} - Q_{ij})(P_{mk} - Q_{mk})]$. Because the sites are independent, when $i \neq m$, $E[(P_{ij} - Q_{ij})(P_{mk} - Q_{mk})] = 0$. Therefore $E[(P_{j} - Q_{j})(P_{k} - Q_{k})] = \sum_{i=1}^{N} E(0) = 0$.

LITERATURE CITED


———. Necessary conditions for the method of inferring phylogeny by linear invariants. (submitted).


BARRY G. HALL, reviewing editor

Received May 29, 1990; revision received July 8, 1990

Accepted August 20, 1990
Letters to the Editor

_Drosophila_ Glucose Dehydrogenase and Yeast Alcohol Oxidase Are Homologous and Share N-terminal Homology with Other Flavoenzymes

Douglas R. Cavener and Philip A. Krasney
Department of Molecular Biology, Vanderbilt University

The presence of similar nucleotide cofactor-binding domains in otherwise unrelated enzymes has been rationalized on the basis of the exon-shuffling model of protein evolution (Gilbert 1978). In the common case of enzymes which share a similar nucleotide-binding domain but exhibit different specificities for primary substrates, the degree of sequence similarity is restricted to small regions specifically demonstrated to bind to the nucleotide cofactor. This pattern of restricted similarity suggests that dissimilar regions arose from nonhomologous exons. We show herein that the sequence similarity of two unrelated flavoenzymes, _Drosophila_ glucose dehydrogenase (GLD) and yeast alcohol/methanol oxidase (AOX/MOX) are unexpectedly similar throughout their primary sequence. In addition these enzymes share considerable similarity both with other unrelated flavoenzymes and with the ras protein for a small amino-terminal region encoding their nucleotide-binding sites.

The genes encoding GLD (E.C. 1.1.99.10) and AOX/MOX (E.C. 1.1.3.13) recently have been isolated from fruit flies (Cavener et al. 1986a; Krasney et al. 1990) and yeast (Ledeboer et al. 1985; Koutz et al. 1989), respectively. The biochemical properties of _Drosophila_ GLD are very similar to those of _Aspergillus oryzae_ GLD, which catalyzes the oxidation of D-glucose to D-glucono-δ-lactone. Unlike the ubiquitous glucose oxidase and AOX/MOX, GLD does not reduce molecular oxygen to hydrogen peroxide (Bak 1967). _Drosophila_ GLD is essential for cuticular modification during development and is secreted in adult males as part of the seminal fluid transferred to females (Cavener and MacIntyre 1983; Cavener et al. 1986b). AOX/MOX is associated with the ability of four yeast genera to utilize methanol as the sole carbon source (Lee and Komagata 1980). It catalyzes the oxidation/reduction of methanol and oxygen to formaldehyde and hydrogen peroxide, respectively. Thus, GLD and AOX/MOX exhibit remarkably different biochemical and functional properties.

A search of both the National Biomedical Research Foundation (NBRF) protein data base and the GenBank nucleic acid data base unexpectedly revealed a significant degree of similarity between _D. melanogaster_ GLD and _Hansenula polymorpha_ (yeast) MOX and _Pichia pastoris_ (yeast) AOX1. The statistical significance was evaluated by comparing the similarity of the primary amino acid sequence of GLD with 200 randomized sequences of AOX1 and MOX by using the RDF2 program of W. Pearson (University of Virginia). The randomized sequences were of the same length and amino acid composition as AOX1 and MOX. The optimized alignment scores for the comparison of AOX1 with GLD and for the comparison of MOX with GLD were greater than 30 SD above the mean of the randomized scores and are more than threefold higher than those of the highest random sequence. Percentage similarities

1. Key words: exon shuffling, protein evolution, ras proto-oncogene, nucleotide binding protein.

2. Present address: Department of Molecular Genetics, Smith, Kline, and French Laboratories, King of Prussia, Pennsylvania 19406.

Address for correspondence and reprints: Douglas R. Cavener, Department of Molecular Biology, Vanderbilt University, Nashville, Tennessee 37235.
between GLD, MOX, and AOX1 were estimated as the number of identical amino acids divided by the average number of residues in the two proteins being compared. The overall sequence similarities are 26% for GLD and AOX1, 26% for GLD and MOX, and 76% for AOX1 and MOX. Alignment of the amino-terminal half of GLD with AOX and MOX requires 26 small (i.e., fewer than six amino acids) gaps, while the carboxyl-terminal half requires two substantial (14- and 51-amino-acid) gaps in addition to 10 small gaps (fig. 1). The precise boundaries of the larger gaps are uncertain because the sequence similarity immediately flanking them is low. The inferred amino acid sequence of GLD contains an additional 57 amino acids in its amino terminus that are not present in AOX1 and MOX. This region contains a hydrophobic domain which may serve as a signal sequence for GLD secretion and which is encoded by a separate exon (Krasney et al. 1990). Although AOX1 and MOX are transported to peroxisomes, they do not contain obvious signal sequences in their amino termini, as expected (Ledeboer et al. 1985 and Koutz et al. 1989). Nonetheless, these seemingly unrelated enzymes share sequence similarity extending throughout almost their entire length. These data are consistent with the hypothesis that GLD and AOX1/MOX diverged from a common ancestral sequence.

Comparing the highly conserved amino-terminal region of GLD, AOX1, and MOX with those of several other flavoproteins revealed a striking similarity within a 28–32-residue region (fig. 2). This region was previously noted to be highly conserved among three types of proteins: p-hydroxybenzoate hydroxylase (PHBH) (Rice et al. 1984), and two related disulfide oxidoreductases, lipoamide dehydrogenase (LPDH) and glutathione reductase (GR). Structural analyses of GR and LDPH demonstrated that this region corresponds to a bαβ segment of the FAD-binding domain and is similar in structure to the Rossmann fold of other nucleotide-binding domains (Thieme et al. 1981; Wierenga et al. 1983). The N-terminal bαβ folds of several proteins that bind NAD show considerable similarity with the FAD-binding domain of human and Escherichia coli GR (Wierenga et al. 1986). Comparing the GLD, AOX1, and MOX sequences with the additional GR and LDPH sequences from other species yields a refined consensus sequence for the FAD-specific ADP-binding bαβ fold (fig. 2). The similarity among the flavoenzymes begins with either Asp or Gln (position 1) followed by four hydrophobic residues (positions 2–5). The canonical glycine triad of all nucleotide bαβ folds exist at positions 6, 8, and 11. The rest of the α helix also exhibits considerable conservation (e.g., Ala, position 15; Arg or Lys, position 17). The second β strand is typically composed of four hydrophobic residues preceding a perfectly conserved Glu. Schulz et al. (1982) have determined that the highly conserved Gly residues and the Glu residue are in close contact with pyrophosphate and ribose moieties of FAD in human glutathione reductase. The GR sequences in the first half of this domain are more similar to GLD, AOX, and MOX than they are to the LDPH sequences. The similarities between the two GR, the GLD, the AOX, and the MOX sequences begin with Phe/Tyr (one residue before position 1), and all five proteins contain GlyGlyGlySer at the bα-α transition, whereas LDPH and PHBH enzymes contain GlyXGlyPro in this region.

Using the Pearson method for statistically evaluating sequence alignments (see above), we failed to detect, beyond the conserved Glu residue at the end of the amino-terminal FAD-binding domain, significant similarity between the two groups of flavoenzymes (GLD/AOX/MOX and GR/LDPH/PHBH). Thus, we believe that this highly conserved FAD-binding domain may have been encoded by a single exon which was integrated into the structure of a variety of functionally unrelated flavoenzymes. This hypothesis is supported by the fact that this conserved domain corresponds closely to a single exon of Drosophila GLD (Whetten et al. 1988; Krasney et al. 1990). Exon III of GLD begins seven codons before the sequence shown in figure 2 and ends with the Ala codon immediately following the conserved Glu. The AOX1 and MOX

Letters to the Editor 145
Fig. 1.—Alignment of *Pichia pastoris* AOX, *Drosophila melanogaster* GLD, and *Hansenula polymorpha* MOX. Asterisks indicate amino acid identities.
yeast genes are completely devoid of introns (Ledeboer et al. 1985; Koutz et al. 1989),
typical of many yeast genes which have apparently lost introns at a rapid rate during
evolution (Fink 1987).

GR and LPDH share several biochemical properties and exhibit 25%-30% se-
quence similarity throughout their entire polypeptide chains (Thieme et al. 1981),
close to the similarity observed between GLD and AOX1/MOX. If these proteins
were to have experienced nearly equal rates of evolution, it would suggest that GR,
LPDH, GLD, and AOX/MOX all may have been constructed by exon shuffling at
nearly the same time. GR and LPDH utilize related substrates, consistent with their
global similarity. Except for the FAD-binding domain, the evolutionary relationship
of GLD and AOX/MOX is much more difficult to rationalize, given their very different
catalytic properties. The similarity of GLD and AOX/MOX sequences in regions which are presumably interacting with very different substrate molecules, suggest considerable evolutionary flexibility in designing substrate specificity. Regardless of the precise molecular events underlying these flavoenzyme’s evolution, they clearly share an ancient kinship through a segment of their FAD-binding domain, a kinship which most likely predates the emergence of prokaryotes and eukaryotes.

A search was conducted in the NBRF and GenBank data bases by using the FAD-binding consensus derived just from GLD, MOX, AOX1, and the two GR sequences. In addition to the flavoenzymes listed in figure 2, ras proteins from a variety of eukaryotes were identified. The three ras proteins exhibit remarkable similarity with the flavoenzymes and display the invariant Gly and Glu residues. This is an important region of the ras protein, since the second Gly residue (residue 12) is the most common site at which spontaneous mutations result in ras-mediated oncogenesis (Barbacid 1987). This glycine is also thought to interact directly with the phosphor-
ibosyl moiety of GDP/GTP (McCormick et al. 1985), perhaps in a manner identical to that observed for the phosphoribosyl moiety of FAD in human GR (Schulz et al. 1982). Nucleotide-binding domains were previously noted to have similar structures, and their evolutionary relationship has been a topic of considerable interest (Ohlsson, et al. 1974; Rossmann and Liljaz 1974; Rossmann and Argos 1977). Most authors have concluded that the similarity among the various subclasses of mononucleotide- and dinucleotide-binding domains is most likely a consequence of their shared evolutionary ancestry, rather than a consequence of their convergence. The data summarized herein give further support to that hypothesis. The degree of primary sequence similarity between the GDP/GTP-binding domain of ras and the FAD-binding domain of GLD, AOX 1, and MOX is considerably more impressive than that usually observed between subclasses of nucleotide-binding domains. Furthermore, like GLD, this region of the human H-ras gene is encoded by a single exon ending with the highly conserved Glu residue (Capon et al. 1983). These facts support a hypothesis that the nucleotide-binding domains of ras and flavoenzymes are similar because they derive from a homologous exon.

LITERATURE CITED


BARRY G. HALL, reviewing editor

Received May 30, 1990; revision received August 13, 1990

Accepted August 13, 1990
Studies on the Phylogenetic Position of the Ctenodactylidae (Rodentia)

Jaap J. Beintema,* Karin Rodewald,† Gerhard Braunitzer,‡
John Czelusniak,† and Morris Goodman‡

*Biochemisch Laboratorium, Rijksuniversiteit Groningen; †Max-Planck-Institut für Biochemie, Abt. Proteinchemie; and ‡Department of Anatomy and Cell Biology, Wayne State University

More than half of all mammalian species belong to the order of the rodents. Although many phylogenetic relationships among rodents have been established beyond doubt, others are not yet established. The phylogenetic position of the Ctenodactylidae (gundis) was discussed in several contributions to a symposium in Paris in 1984 (Luckett and Hartenberger 1985). This is a small rodent family with only four species living in the border regions of the Sahara Desert in Africa. This taxon has been grouped either with the hystricognathous rodents or as a separate branch, as a very early offshoot from the rodent stem: Luckett (1985), Bugge (1985), and George (1985) concluded, from different sets of data, that the Ctenodactylidae share with the hystricognathous rodents a number of derived traits and that the Ctenodactylidae should be considered as an early offshoot of the latter. Wood (1985) discussed the phylogenetic position of several sciurognathous hystricomorph families, including the Ctenodactylidae. These families are also restricted to Africa and have often been associated with the hystricognathous rodents. However, in Wood's opinion there are no indications for these relationships, the similarities being the result of conservation of ancestral features or of convergence. Hartenberger (1985) concluded from the fossil evidence that the Ctenodactylidae are one of the oldest recognized rodent families, diverging from all other rodent taxa in the Lower Eocene. With additional molecular data, it may be possible to shed more light on this controversy and to learn more about the origin of the rodents.

No molecular evidence is yet available on the phylogenetic position of the Ctenodactylidae. Therefore, several tissues and blood of gundis were collected, and a number of proteins were isolated and investigated.

Gundi myoglobin was isolated, and its amino acid sequence was determined (Beintema et al. 1990). Only four other rodent myoglobin sequences have been determined so far: mouse (Mus musculus) (Harris et al. 1985), casiragua (Proechimys guairae) (Harris et al. 1985), viscacha (Lagostomus maximus) (Gurnett et al. 1984), and mole rat (Spalax ehrenbergi) (Gurnett et al. 1984). The gundi amino acid myoglobin sequence was compared with the 73 other known vertebrate myoglobin amino acid sequences by using computer procedures, described by Czelusniak et al. (1990), that search for the tree or set of trees with shortest nucleotide substitution length. The shortest trees for the 74 myoglobin sequences found in this search failed to represent rodents, primates, and artiodactyls as monophyletic orders. In these shortest trees, the gundi joins the hystricognath (casiragua/viscacha) branch, which is rather widely separated from the mouse/mole rat branch. If the four other rodents are forced together, which requires four additional nucleotide substitutions, the branch to the gundi still joins the two hystricognaths, or, for the same score, the gundi branches off before the divergence of the other four rodent sequences.

The amino acid sequences of the α and β hemoglobin chains of gundis were
determined (Beintema et al. 1990) and compared with the α and β hemoglobin sequences from 134 vertebrate species—α sequences by themselves, β chains by themselves, and tandemly combined α and β chains. Parsimony procedures were used to find the shortest tree. When the β sequences are used, the gundi groups with the tree shrew, and the common branch of those two joins lemur and then joins a branch with nine rodents and pig. When the α sequences are used, gundi groups with squirrel, but this branch does not join the other rodents. However, when tandemly combined α and β sequences are used, a monophyletic tree of the rodents is obtained, with the first intrarodent divergence separating gundi from all the other rodents (fig. 1). Joining gundi with the hystricognath guinea pig results in a tree with two more substitutions. This is a low number compared with the branch lengths connecting gundi and guinea pig to the rest of the rodent stem.

Studies of other proteins provide little additional evidence about the phylogenetic position of the Ctenodactylidae. Gundi αA-crystallin shares a replacement with two hystricognath species that separates these taxa from other rodents (Hendriks et al. 1987). Immunological studies of serum proteins indicate no special affinity between

![Diagram of the most parsimonious tree of tandemly combined α and β sequences of rodent hemoglobins.](image)

gundi and any other investigated rodent species (Sarich 1985, and personal communication).

The molecular data summarized in this letter confirm that both the Ctenodactylidae and the hystricognathous rodents are early offshoots in the order of the rodents, but—as in classical phylogeny—it is impossible yet to decide whether they share a common ancestor or are located on separate branches.

**Sequence Availability**

The amino acid sequences of gundi myoglobin, α hemoglobin chain, and β hemoglobin chain have been deposited in the NBER Data Library under accession numbers A33082, B33082, and C33082, respectively.

**Acknowledgments**

Gundis were caught by Said Chouiref (Beni Barka, Foum Tataouine, Tunisia), with cooperation of the Service d’Agriculture in Tataouine and with permission from the Ministère de l’Agriculture (Direction des Forêts) of Tunisia. We thank Dr. Kenneth A. Josey (Cambridge, United Kingdom) for financial support of the trip to Tunisia and for his interest in these studies, and we thank Dr. R. N. Campagne for carefully reading the manuscript.

**LITERATURE CITED**


Hartenberger, J. L. 1985. The order Rodentia: major questions on their evolutionary origin,


Walter M. Fitch, reviewing editor

Received May 14, 1990; revision received August 7, 1990

Accepted August 7, 1990
For the style of citations, please note the following examples:

Journal articles:


Books:


Book chapters:


The abbreviations of periodicals should be those used in Chemical Abstracts.

Articles should include “Reviewed under the auspices of (name of the editor or associate editor with whom author has corresponded)” at the end of the Literature Cited section.

Footnotes.—Footnotes should be used sparingly. When necessary, they should be indicated in the text by superscript arabic numerals; the notes themselves should be typed on a page separate from the text. Footnotes to tables are referenced by superscript letters, except for significance levels, which use asterisks; table footnotes should be typed on the same page as the table to which they pertain.

Tables.—Each table must have a brief and self-explanatory title, be numbered with arabic numerals in order of its appearance in the text, and be typed on a separate page. Large, complex tables are discouraged. Guidelines for table format may be found in the CBE Style Manual and the Chicago Manual of Style or may be obtained by writing the editors of this journal.

Legends.—Figure legends should be typed on pages at the end of the manuscript, after tables. Each legend must be descriptive so that the illustration can be understood apart from the text and must define abbreviations used in the illustration.

Illustrations.—Each illustration (figure) should be an original, not a photocopy. Illustrations should be separate and have uniform lettering. They should be numbered consecutively, following the sequence in which they are mentioned in the text. The place where each illustration is to be inserted may be indicated by a circled note in the margin of the typescript. Names of authors, figure number, and an arrow indicating proper orientation should be written lightly in pencil on the back of each figure. Line drawings must be of high quality; typewritten or hand lettering is unacceptable. Photographs should be high-contrast, glossy prints. Magnifications may be indicated by a micron bar or in the legend. Photographs to be reproduced without further reduction must be so marked and may not exceed 4¾ inches wide by 7¾ inches long (122 by 194 mm), in order to fit the journal format. Please keep in mind that these dimensions are maxima. Because of the need for a figure legend, the illustration cannot be the maximum size in both dimensions.

Proofs and Reprints

Offprint order forms will be sent to the author (or, in the case of multiple authors, to the senior author) with page proofs. There will be no page charge for publication.
Information for Contributors

Molecular Biology and Evolution is a bimonthly journal devoted to the interdisciplinary science between molecular biology and evolutionary biology. The journal will emphasize experimental papers, but theoretical papers are also published if they have a solid biological basis. Although this journal is primarily for original papers, review articles and book reviews are also published. Brief discussion and comment on material published in this journal or on issues particularly relevant to readers of this journal will be published as “Letters to the Editor.” Letters that refer to a paper handled by an associate editor should be sent to that editor or to the editor in chief.

To minimize publication delays, authors should follow the instructions given here and should also provide their telephone numbers.

Submission of Manuscripts

Send manuscripts (one original and two high-quality copies) to the editor in chief, managing editor, or any associate editor. Any manuscript or any part of a manuscript that has been published or submitted for publication elsewhere cannot be accepted for publication. Anyone who wishes to write a review article should contact the managing editor. Correspondence about book reviews should also be directed to the managing editor. Decision on acceptance of papers will be made as rapidly as possible. Papers that are not suitable to the journal will be returned immediately to authors without detailed review.

After a manuscript is accepted, its author will be requested to sign an agreement transferring copyright to the publisher. No published material may be reproduced or published elsewhere without the written permission of the copyright owner. The journal will not be responsible for the loss of manuscripts at any time.

Publication is taken imply that the authors are prepared to make available to the public any clone of cells, DNA, or antibodies used in reported experiments and any computer programs or nucleic acid or protein sequences on which the paper is based. Nucleic acid sequences must be provided to GenBank in computer-readable format along with the appropriate associated data. The GenBank accession number will be printed in the article.

Preparation of Manuscripts

Papers must be written in English and organized in the sequence described below. Each section must be typed double-spaced on heavyweight, nonerasable bond; the page margins should be 1½ inches wide to allow for corrections, manuscript editor’s notes and queries, and typesetting markup. Type should not be more than 12 characters per inch nor more than six lines per inch. Special typefaces (e.g., italic or sans serif) should not be used, and right-hand margins should not be justified. Word-processing output on dot matrix printers is acceptable only if it is the quality of the standard typewriter. Handwritten items (e.g., Greek letters) must be identified in the margin. Non-English words must have correct diacritics. Although each major part of the paper (e.g., Literature Cited) must begin on a new page, the pages should be numbered consecutively throughout, beginning with the title page and continuing through the abstract, text, Appendix, Literature Cited, footnotes, tables, and ending with figure legends.

Title page.—This page should contain the paper’s title, the names of all authors, the institution(s) at which the research was done, the current affiliations of all authors, the name and address for correspondence, and a footnote on nonstandard abbreviations used, if any (see below). Finally, the title page should also provide a running head (maximum of 50 characters and spaces).

Abstract.—The abstract should be a one-manuscript-page factual condensation of the entire paper, including a statement of purpose, a clear description of observations and findings, and a concise presentation of conclusions. It should not assert that the findings are discussed.

Key words.—A list of three to six words or phrases should be provided that will accurately index the subject matter of the article.
Text.—The text should comprise the following sections: (1) Introduction, (2) Material and Methods, (3) Results, (4) Discussion, and (5) Acknowledgments (if any). Papers should be concise but will not be restricted in length.

All organisms mentioned must be identified by their scientific binomens and underlined. Symbols for genetic loci must also be underlined and should follow the established rules of genetic nomenclature for the various organisms (consult M. Demerec et al., Genetics 54:61–76, 1966). Include the formal IUB name and number of all enzymes mentioned.

Do not use abbreviations for words or phrases used less than five times. Abbreviations used by the Journal of Biological Chemistry will be regarded as standard; nonstandard abbreviations should be defined collectively in a footnote.

Mathematical equations must be carefully typewritten: spacing between characters should be correct as typed. It will be assumed that all characters in equations and their counterparts in the text will be set in italics unless the author specifies otherwise the first time a character appears. Equations should be numbered sequentially, in arabic numerals in parentheses, on the right-hand side of the page.

These and other guidelines can be found in the Council of Biology Editors Style Manual (4th ed., 1978). In general, all material should conform to the CBE format. See also recent issues of this journal.

Terminology.—A satisfactory interdisciplinary communication requires using words with careful attention to their precise meaning in both disciplines. Authors may use any word they choose provided only that its meaning is clear, consistent, and serves to increase the paper’s comprehensibility. The following preferred usages are not prescriptive but will be assumed unless authors define them otherwise.

Where the alignments disagree, they are differences rather than changes, since there may have been multiple changes to create a single difference. Differences or changes are replacements if the sequences are amino acids, substitutions if they are nucleotides. Mutations should be restricted to changes before selection has operated. Homology must be defined, since it has two common meanings: (1) observed similarity and (2) inferred common ancestry. The term similarity is preferred for meaning 1 because sequences may have similarity acquired by convergence (analogy) rather than retained after divergence (homology). When homology arises via a gene duplication (all or part), it is properly called paralogy; when it arises via speciation, it is properly called orthology. Gaps are introduced into sequences to increase their similarity rather than to optimize similarity (homology), unless an algorithm is employed that guarantees an optimized result according to the way similarity (homology) is defined (e.g., as maximum matches—a third meaning of homology). Similarity should not be asserted to be significant unless patently obvious or accompanied by a probability statement and its method of determination ($\chi^2$, standard measure, binomial, etc.).

The preferred single-letter code for nucleotide bases including ambiguity is: A = adenine, C = cytosine, G = guanine, T = thymine, U = uracil, R = A/G (purine), Y = C/T (pyrimidine), M = A/C, W = A/T, S = C/G, K = G/T, B = C/G/T (not A), D = A/G/T (not C), H = A/C/T (not G), V = A/C/G (not T), N = X = A/C/G/T (any or unknown). For ambiguous nucleotides, T and U are equivalent.

Literature Cited.—Literature in the text should be cited by author and year and, where citation is to a book, the relevant pages thereof. Text citations of two or more works at a time should be given in chronological order. When a paper written by three or more authors is cited, write the name of the first author plus “et al.” The Literature Cited section at the end of the paper should be arranged alphabetically by author(s) and then chronologically and should contain only works specifically cited in the text. References to papers that have not yet been published will be as for articles (see below), except that “accepted” (along with the journal name) will replace the volume and page numbers. “In press” will not be used. (When such papers include authors of the submitted manuscript, copies of those papers must accompany the submitted manuscript.)
STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION
(Required by 39 U.S.C. 3685)

1. Title of publication: Molecular Biology and Evolution
   a. Publication number: 07374038
2. Date of filing: October 24, 1990
3. Frequency of issue: Bi-Monthly
   a. No. of issues published annually: 6
   b. Annual subscription price: $190.00
4. Location of known office of publication: 5720 S. Woodlawn Avenue, Chicago, Illinois 60637.
5. Location of the headquarters or general business offices of the publishers: 5720 S. Woodlawn Avenue, Chicago, Illinois 60637.
6. Names and addresses of publisher, editor, and managing editor:
   Publisher: The University of Chicago Press, 5720 S. Woodlawn Avenue, Chicago, Illinois 60637
   Editor: Walter M. Fitch, University of California Irvine, Irvine, California 92717
   Managing Editor: Masatoshi Nei, University of Texas, Health Science Center, Houston, Texas 77225
8. Known bondholders, mortgagees, and other security holders owning or holding 1 percent or more of total amount of bonds, mortgages or other securities: None
9. The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes have not changed during preceding 12 months.

10. Extent and nature of circulation:
    | Average Number of Copies Each Issue During Preceding 12 Months | Single Issue Nearest To Filing Date |
    |---------------------------------------------------------------|-----------------------------------|
    | A. Total number copies printed                               | 1,200                             |
    | B. Paid circulation:                                         |                                    |
    | 1. Sales through dealers and carriers, street vendors and counter sales | 19                               |
    | 2. Mail subscriptions                                       | 800                               |
    | C. Total paid circulation                                    | 819                               |
    | D. Free distribution:                                        | 54                                |
    | 1. Office use, leftover, unaccounted, spoiled after printing | 127                               |
    | 2. Returns from news agents                                  | 275                               |
    | E. Total distribution (sum of C & D)                         | 873                               |
    | F. Copies not distributed:                                  | 925                               |
    | G. Total (sum of E and F)                                    | 1,200                             |

I certify that the statements made by me above are correct and complete.

ROBERT SHIRRELL, Journal Manager

This publication is available in microform from UMI.

Please send me information about the titles I’ve listed below:

________________________________________________________

Name__________________________
Title__________________________
Company/Institution______________
Address________________________
City/State/Zip__________
Phone (____-______)__________

UMI
A Bell & Howell Company
300 North Zeeb Road, Ann Arbor, MI 48106 USA
800-521-0600 toll-free
313-761-4700 collect from Alaska and Michigan
800-343-5298 toll-free from Canada
Forthcoming

Strong Reproductive Isolation between Closely Related Tropical Sea Urchins (Genus Echinometra)
Stephen R. Palumbi and Edward C. Metz
5S Ribosomal RNA Sequence Inappropriate for Phylogenetic Reconstruction
Kenneth M. Halanych
Assessing the Reliability of 5S rRNA Sequence Data for Phylogenetic Analysis in Green Plants
Kelly P. Steele, Kent F. Holsinger, Robert K. Jansen, and David W. Taylor
Distribution of hobo Transposable Elements in Natural Populations of Drosophila melanogaster
L. Pascual and G. Periquet
Isolation and Molecular Evolutionary Analysis of a Cytochrome c Gene from Oryza sativa (Rice)
Elizabeth C. Kemmerer, Ming Lei, and Ray Wu
Molecular Evolution of the Ψη-Globin Gene Locus: Gibbon Phylogeny and the Hominoid Slowdown
Wendy J. Bailey, David H. A. Fitch, Danilo A. Tagle, John Czelusniak, Jerry L. Slightom, and Morris Goodman
Evolutionary Origin and Diversification of the Mammalian CD1 Antigen Genes
Austin L. Hughes
What Is a Genus in the Trypanosomatidae Family? Phylogenetic Analysis of Two Small rRNA Sequences (Letter to the Editor)
Erika Gómez, Ana Mariá Valdés, Daniel Piñero, and Roberto Hernández
Study of an Isolated Population at the Nucleotide Level: rp49 Region of a Canarian Population of Drosophila subobscura
Julio Rozas and Monteserrat Aguadé
Original Articles

1 Subfamily Relationships and Clustering of Rabbit C Repeats
   Dan E. Krane, Andrew G. Clark, Jan-Fang Cheng, and Ross C. Hardison

31 Comparison of Alcohol Dehydrogenase Expression in Drosophila melanogaster and D. simulans
   M. Scott Thomson, James W. Jacobson, and Cathy C. Laurie

49 Rates of DNA Change and Phylogeny from the DNA Sequences of the Alcohol Dehydrogenase Gene for Five Closely Related Species of Hawaiian Drosophila
   Robert G. Rowan and John A. Hunt

71 Ribosomal DNA Variation Within and Between Species of Rodents, with Emphasis on the Genus Onychomys
   Marc W. Allard and Rodney L. Honeycutt

85 Variation in Mitochondrial Cytochrome b Sequence in Natural Populations of South American Akodontine Rodents (Muridae: Sigmodontinae)
   Margaret F. Smith and James L. Patton

104 Mitochondrial DNA Polymorphisms in the Two Subspecies of Drosophila sulfurigaster: Relationship between Geographic Structure of Population and Nucleotide Diversity
   Koichiro Tamura, Tadashi Aotsuka, and Osamu Kitagawa

115 The Relative Rate of DNA Evolution in Primates
   Simon Easteal

128 Methods for Inferring Phylogenies from Nucleic Acid Sequence Data by Using Maximum Likelihood and Linear Invariants
   W. C. Navidi, G. A. Churchill, and A. von Haeseler

Letters to the Editor

144 Drosophila Glucose Dehydrogenase and Yeast Alcohol Oxidase Are Homologous and Share N-terminal Homology with Other Flavoenzymes
   Douglas R. Cavener and Philip A. Krasney

151 Studies on the Phylogenetic Position of the Ctenodactylidae (Rodentia)
   Jaap J. Beintema, Karin Rodewald, Gerhard Braunitzer, John Czelusniak, and Morris Goodman
Molecular Biology and Evolution (ISSN 0737-4038) is published six times a year in January, March, May, July, September, and November by the University of Chicago Press, 5720 South Woodlawn Avenue, Chicago, Illinois 60637, and sponsored by the Molecular Biology and Evolution Society, the Society for the Study of Evolution, and the American Society of Naturalists. Molecular Biology and Evolution is indexed in Current Contents/Life Sciences, Science Citation Index, and Excerpta Medica. Advertising space in Molecular Biology and Evolution is available, as is rental of its subscriber list. For information and rates, please contact the advertising sales staff, The University of Chicago Press, Journals Division, 5720 South Woodlawn Avenue, Chicago, IL 60637; phone (312) 702-8187 or (312) 702-7361. Advertising and list rental is limited to material of scholarly interest to our subscribers.

Subscription Rates U.S.A.: institutions, 1 year $190.00; individuals, 1 year $47.00. Special ASN and SSE individual members, 1 year $37.00. Student subscription rate, U.S.A., 1 year $35.00 (copy of student ID must accompany subscription). Other countries add $6.00 for each year’s subscription to cover postage. For airmail delivery, the rate is $11.00 for Canada and Mexico and $16.00 elsewhere. Subscription agent for Japan: Kinokuniya Company, Ltd. Single copy rates: institutions $31.75, individuals $8.00. Business correspondence should be addressed to The University of Chicago Press, Journals Division, P.O. Box 37005, Chicago, Illinois 60637.

Change of Address Please notify the Press and your local postmaster immediately, giving both your old and new addresses. Allow four weeks for the change. Postmaster: Send address changes to Molecular Biology and Evolution, The University of Chicago Press, P.O. Box 37005, Chicago, Illinois 60637. Claims for missing numbers should be made within the month following the regular month of publication.

Editorial Correspondence Correspondence about book reviews should be addressed to the Managing Editor: Masatoshi Nei, Director of the Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, Pennsylvania 16802. Editorial correspondence should be addressed to the Editor in Chief: Walter M. Fitch, Department of Biological Sciences, University of Southern California, University Park, Los Angeles, California 90089-1340. For manuscript submission, see Information for Contributors at the end of this issue.

Copying beyond Fair Use The code on the first page of an article in this journal indicates the copyright owner's consent that copies of the article may be made beyond those permitted by Sections 107 or 108 of the U.S. Copyright Law provided that copies are made only for personal or internal use, or for the personal or internal use of specific clients, and provided that the copier pay the stated per copy fee through the Copyright Clearance Center, Inc., 27 Congress Street, Salem, Massachusetts 01970. To request permission for other kinds of copying, such as copying for general distribution, for advertising or promotional purposes, for creating new collective works, or for resale, kindly write to Permissions Department, The University of Chicago Press, 5801 South Ellis Avenue, Chicago, Illinois 60637.

Second-class postage paid at Chicago, Illinois, and at additional mailing office.

© 1991 by The University of Chicago. All rights reserved.