Estimating Divergence Times of *Drosophila* Species from DNA Sequence Comparisons

Simon Easteal and John G. Oakeshott
Department of Population Biology, Australian National University

Comparisons between species of the DNA sequences of homologous single copy genes have great potential for use in the reconstruction of phylogenies. The rates of change of nucleotides in introns and "silent" sites in exons are independent of generation time, approximately constant in different lineages, and apparently also constant for different genes (Hayashida and Miyata 1983). From calibration with the fossil record, Hayashida and Miyata (1983) have estimated the rates ($v$) to be $3.7 \times 10^{-9}, 5.3 \times 10^{-9},$ and $5.6 \times 10^{-9}$ changes per site per year for short (<300 bp) introns, long (>500 bp) introns, and exon silent sites, respectively. These rates can now be used in comparative DNA sequence analyses to estimate divergence times ($t$) among other taxa lacking fossil records.

This approach was recently used by Bodmer and Ashburner (1984), who described the nucleotide sequences of the Alcohol dehydrogenase (*Adh*) genes of four members of the *Drosophila melanogaster* species group and, confining their analysis to exon silent sites, inferred divergence times among these species from comparison of their sequences. They estimated that *D. melanogaster* diverged from *D. orena* 13 Myr ago and from *D. simulans* 3.9 Myr ago, while *D. simulans* and *D. mauritiana* separated 2.9 Myr ago. Similar results were obtained by Cohn et al. (1984), who compared partial *Adh* sequences among three of these species and estimated that *D. simulans* and *D. mauritiana* diverged from each other 2.7 Myr ago and from *D. melanogaster* 4.5 Myr ago. However, Cohn et al. (1984) used a different rate constant, applicable to melting profiles of whole genome single copy DNA (Zwiebel et al. 1982).

Since whole genome single copy DNA comprises various categories of DNA evolving at very different and not necessarily constant rates, the application of a rate constant derived from whole genome single copy DNA to any specific sequences is of doubtful validity. We advocate use of the constants obtained by Hayashida and Miyata (1983) because they are restricted to categories of DNA shown to evolve at constant rates and they are calculated separately for categories evolving at different, constant rates. Nevertheless, there remain important problems in the use of the Hayashida and Miyata constants. Here we discuss these problems around an extended analysis of Bodmer and Ashburner's *Adh* data and a parallel analysis of the partial sequences of the Heat Shock Cognate (*Hsc*) gene of *D. melanogaster* and *D. simulans* (Ingolia and Craig 1982).

We used published sequence alignments, which were unambiguous, to estimate the number of nucleotide substitutions per site ($K'$) in each *Adh* region defined by Bodmer and Ashburner for all species pairs, and in exon I and intron I of *Hsc* for *D. melanogaster* and *D. simulans*. We used criteria of Hayashida and Miyata (1983) to determine effective lengths of introns and effective numbers of silent sites in exons and to correct for multiple substitutions.

1. Key words: *Drosophila, Adh, Hsc*, DNA sequence comparisons, divergence times.

Address for correspondence and reprints: John G. Oakeshott, Department of Population Biology, Research School of Biological Sciences, Australian National University, P.O. Box 475, Canberra City ACT 2601, Australia.

© 1985 by The University of Chicago. All rights reserved.
0737-4038/02/0200-0212$02.00

87
Table 1 gives $K^c$ values among the four species for $Adh$ regions and between $D. melanogaster$ and $D. simulans$ for $Hsc$ regions. As expected, $K^c$ values are heterogeneous ($F[13, 50] = 20.0, P < 0.001$ on square root transforms) across regions, and most (69%) of this heterogeneity is among the more divergent regions of the genes (i.e., the introns and exon silent sites used by Hayashida and Miyata in determining their rate constants) and the remaining, comparatively conserved regions. Importantly, however, there is still highly significant heterogeneity ($F[7, 50] = 8.6, P < 0.001$) among the divergent regions, and relatively little (15%) of the latter heterogeneity lies among the three categories of divergent region recognized by Hayashida and Miyata (short and long introns and silent exon sites). Estimates of $t$ derived from $K^c$ values for the divergent regions of $Adh$ and $Hsc$ (table 2) are therefore also heterogeneous ($F[7, 25] = 8.5, P < 0.001$).

A negligible proportion (<1%, $F[1, 25] = 0.2, P > 0.05$) of the heterogeneity in $t$-values arises from overall differences between $Adh$ and $Hsc$. This supports the conclusion of Hayashida and Miyata that the rate of the DNA "clock" is similar for different genes. It is also noteworthy that only 4% of the heterogeneity lies between the exon silent sites and the introns ($F[1, 25] = 2.7, P > 0.05$); this suggests that the strong bias in codon usage among the $Drosophila$ genes noted by Bodmer and Ashburner (1984) did not introduce major errors into the estimates from the exon silent sites.

We are therefore left with a serious problem: threefold differences in $t$ estimates are obtained from different regions which a priori are eligible for inclusion in the analysis. Most of the heterogeneity (86%) lies between two arbitrary sets of regions, one consisting of Intron I and Exons II and III of $Adh$ (which have diverged relatively slowly) and the other consisting of Introns II and III and Exon I of $Adh$ and Intron I and Exon I of $Hsc$ (which have diverged more rapidly). Within each of these sets of regions there is no significant heterogeneity of $t$-values ($F[6, 25] = 1.2, P > 0.05$).

The estimates of $t$ based on the five more rapidly diverging regions are $9.4 \pm 1.2$ Myr for $D. melanogaster$ versus $D. simulans$ and $D. mauritiana$, $7.7 \pm 0.9$ Myr for $D. simulans$ versus $D. mauritiana$, and $30.2 \pm 2.4$ Myr for $D. orena$ versus $D. melanogaster$, $D. simulans$, and $D. mauritiana$. These compare with values of $6.6 \pm 1.1$, $4.8 \pm 1.4$, and $20.1 \pm 2.7$ Myr for the corresponding comparisons using all eight regions, and with values of $3.0 \pm 0.4$, $2.0 \pm 0.6$, and $9.9 \pm 0.6$ for these comparisons using the three more slowly diverging regions. Bodmer and Ashburner's estimates of $3.9, 2.9$, and $13$ Myr resemble the estimates based on the more slowly diverging regions, because theirs are based on two out of three of these regions and on only one of the more rapidly diverging regions.

Since the heterogeneity of $t$ estimates between regions is consistent across the different species comparisons, it cannot be attributed to sampling errors. Which then are the most appropriate estimates? Either otherwise general structural or functional constraints are relaxed in the five relatively rapidly evolving regions, or constraints additional to those operating in other regions exist in the three more slowly evolving regions. The second possibility seems to us the more plausible. There are fewer of the slower-evolving regions, and one, Intron I of $Adh$, lies not between two exons but between the two $Adh$ promoters, where structurally or functionally constrained regulatory sequences might be expected. With regard to the other two more slowly evolving regions, the silent sites in $Adh$ Exons II and III, we note that Hayashida and Miyata found other cases of "unusually strong sequence conservation" of silent sites in exons and excluded these in their calculation of rate constants. We therefore argue that the homogeneous estimates of $t$ based on the five more rapidly evolving regions are the most appropriate.

Strict application of the Hayashida and Miyata methods thus predicts that
Table 1
Nucleotide Substitutions per Site \( (K_c) \) in Different Regions of Adh and Hsc among the Four Species

<table>
<thead>
<tr>
<th></th>
<th>Adh</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>( K_c \times 10^3 )</td>
</tr>
<tr>
<td>Sequence Length (bp)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Divergent regions:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intron I</td>
<td>451–645</td>
<td>35</td>
<td>42</td>
</tr>
<tr>
<td>Exon I (silent)</td>
<td>91</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>Intron II</td>
<td>37–66</td>
<td>63</td>
<td>46</td>
</tr>
<tr>
<td>Exon II (silent)</td>
<td>99.3</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>Intron III</td>
<td>66–76</td>
<td>96</td>
<td>113</td>
</tr>
<tr>
<td>Exon III (silent)</td>
<td>59.3</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td>Conserved regions:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult leader</td>
<td>134</td>
<td>30</td>
<td>23</td>
</tr>
<tr>
<td>Larval leader</td>
<td>59–70</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Exon I (substitutions)</td>
<td>76.3</td>
<td>13</td>
<td>26</td>
</tr>
<tr>
<td>Exon II (substitutions)</td>
<td>305.7</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Exon III (substitutions)</td>
<td>204.7</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

\[\text{NOTE.-}\text{mel = melanogasler, sim = simulans, mau = mauritiana, and ore = orca. The regions of both genes are divided into "divergent" (those used by Hayasida and Miyata in determining their rate constants) and "conserved" categories.}\]
species within the melanogaster subgroup of the melanogaster group are up to about 30 Myr old. This is consistent with Throckmorton's (1975) estimate for the age of the subgroup based on biogeographic considerations. He argued that the melanogaster species group existed over 35 Myr ago and that a radiation within the group, including the melanogaster subgroup, occurred in Africa, possibly starting in the late Oligocene (33–25 Myr ago). Our \( t \) estimate of 9.4 ± 1.2 Myr between \( D. \) melanogaster and \( D. \) simulans also concurs with estimates based on differences in melting profiles of whole genome single copy DNA, \( \Delta t_m \). Estimates of \( \Delta t_m \) between \( D. \) melanogaster and \( D. \) simulans range from 1.9 to 5 (Laird and McCarthy 1968; Zwiebel et al. 1982); using \( \Delta t_m \) rate constants of 0.2–0.45 per Myr determined by Hunt et al. (1981) and Hunt and Carson (1983) for the Hawaiian Drosophila, these provide \( t \) estimates ranging from 2.1 to 12.5 Myr between \( D. \) melanogaster and \( D. \) simulans.

Our final conclusions relate back to the heterogeneity in \( t \) estimates across different regions that a priori were eligible for inclusion in the analysis. This heterogeneity is inconsistent with the premise of Hayashida and Miyata that rates of substitution are relatively constant across different introns or silent sites in different exons. The Hayashida and Miyata methods have enormous potential for use in phylogenetic reconstruction. However, our Drosophila analyses highlight three areas in need of further attention before this potential can be fully exploited. First, there is a clear need for further calibration of evolutionary rates, \( v \), for both introns and exons, between taxa with good fossil records. Second, \( K' \) values for several regions are necessary from taxa without fossil records before the \( v \) constants can be applied meaningfully to obtain estimates of the species' divergence times. Third, we need to formulate more stringent criteria for assessing heterogeneity in \( K' \) values among regions and deciding the eligibility of values from different regions for inclusion in the dating analyses.

Acknowledgments

We thank I. A. Boussy, C. Collet, and N. E. Pierce for invaluable discussions and comments on the manuscript.

LITERATURE CITED

## Table 1
Nucleotide Substitutions per Site ($K^c$) in Different Regions of Adh and Hsc among the Four Species

<table>
<thead>
<tr>
<th></th>
<th>Adh</th>
<th></th>
<th>Hsc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sequence Length (bp)</td>
<td>$K^c \times 10^3$</td>
<td>Sequence Length (bp)</td>
</tr>
<tr>
<td>Divergent regions:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intron I</td>
<td>451–645</td>
<td>35</td>
<td>42</td>
</tr>
<tr>
<td>Exon I (silent)</td>
<td>22.7</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>Intron II</td>
<td>37–66</td>
<td>63</td>
<td>46</td>
</tr>
<tr>
<td>Exon II (silent)</td>
<td>99.3</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>Intron III</td>
<td>66–76</td>
<td>96</td>
<td>113</td>
</tr>
<tr>
<td>Exon III (silent)</td>
<td>59.3</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td>Conserved regions:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult leader</td>
<td>134</td>
<td>30</td>
<td>23</td>
</tr>
<tr>
<td>Larval leader</td>
<td>59–70</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Exon I (substitutions)</td>
<td>76.3</td>
<td>13</td>
<td>26</td>
</tr>
<tr>
<td>Exon II (substitutions)</td>
<td>305.7</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Exon III (substitutions)</td>
<td>204.7</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

**NOTE.**—mel = *melanogaster*, sim = *simulans*, mau = *mauritiana*, and ore = *orena*. The regions of both genes are divided into “divergent” (those used by Hayashida and Miyata in determining their rate constants) and “conserved” categories.


WALTER M. FITCH, reviewing editor

Received August 3, 1984; revision received September 24, 1984.
Biochemical Pathways in Prokaryotes Can Be Traced Backward through Evolutionary Time

Roy A. Jensen
Center for Somatic-Cell Genetics and Biochemistry, Department of Biological Sciences, State University of New York at Binghamton

For the first time, a credible prokaryotic phylogenetic tree is being assembled by Woese and others using quantitative sequence analysis of oligonucleotides in the highly conservative rRNA. This provides an evolutionary scale against which the evolutionary steps that led to the arrangement and regulation of contemporary biochemical pathways can be measured. This paper presents an emerging evolutionary picture of aromatic amino acid biosynthesis within a large superfamily assemblage of prokaryotes that is sufficiently developed to illustrate a new perspective that will be applicable to many other biochemical pathways.

The Emerging Phylogeny of Prokaryotes

Until recently, the phylogenetic past of prokaryotes was generally conceded to be forever lost. The lack of a fossil record, inevitable instances of convergent evolution, and possible scenarios of "horizontal" gene transfer by promiscuous extrachromosomal agents posed seemingly insurmountable barriers. Fortunately, horizontal gene transfer in natural populations seems limited to genes whose functions are usually dispensable in nature (Campbell 1981). The idea (Zuckerkandl and Pauling 1965) that macromolecular sequences might be sufficiently conserved to serve as "documents of evolution"—being, in essence, living fossils—has been realized in oligonucleotide cataloging (Fox et al. 1980; Stackebrandt and Woese 1981), a technique that lends itself to the computer-assisted estimation of an evolutionary similarity between any pair of taxa/organisms. This, expressed as a similarity coefficient (SAB), has been discussed at length by Fox et al. (1977). At one extreme, catalogs obtained from samples of identical cell populations would yield an SAB of 1.0. These catalogs provide a basis for estimating the evolutionary relationships of the organisms from which the catalogs were derived. The resulting new phylogenetic classification often differs drastically from the pragmatic classification offered by Bergey's Manual of Determinative Bacteriology. This is illustrated by the fact that species currently named within a single genus (e.g., Rhodopseudomonas gelatinosa and Rp. capsulata) are now not even considered to belong to the same family (Fox et al. 1980; Stackebrandt and Woese 1981).

Evolution of Biochemical Pathways

Metabolic pathways are a source of more biochemical diversity in nature than is generally appreciated. Not only may pathway enzymes differ in cofactor specificities, metal requirements, allosteric specificities, and multifunctional capabilities, but the

1. Key words: microbial phylogeny, evolution, aromatic biosynthesis, cofactor specificity. Abbreviations: DAHP = 3-deoxy-D-arabino-heptulosonate 7-phosphate; NAD+ = nicotinamide adenine dinucleotide; NADP+ = nicotinamide adenine dinucleotide phosphate.
enzymic-step construction of the biochemical pathways may differ. Organisms sharing the same biochemical state for characters such as those listed above are not necessarily genealogically close. Neither does the recognition of different pathway character states provide a compelling basis for conclusions about pathway evolution. On the other hand, if organisms under study have known positions on a credible phylogenetic tree, then deductions about pathway evolution are feasible. The biochemical-pathway data per se are not used to construct the phylogeny because: (1) no quantitative measure comparable to an $S_{AB}$ value is available; (2) the various character states are not equivalent to one another—compare, for instance, such features as the presence or absence of an enzyme, enzyme sensitivity to inhibition by compound a or by compound b, or the presence or absence of a multifunctional protein (indicating occurrence or nonoccurrence of a gene fusion event); and (3) the impact of a given pathway on overall metabolism differs from organism to organism because biochemical diversity in nature is considerable (e.g., an amino acid may be a precursor of a quantitatively significant pigment in one microbe but not in others). Thus, the various character states of biochemical pathways are not readily translated into a number that expresses evolutionary distance ([1] and [2] above), and a given character state is undoubtedly not equally constrained in different organisms ([3] above). Hence, our approach is to assume the phylogeny established by oligonucleotide cataloging as the basis for reconstruction of the evolutionary history of aromatic amino acid biosynthesis. A consistent, cladistic methodology can then be employed in which the most parsimonious solutions are adopted.

**Diversity of Aromatic Biosynthesis and Regulation**

In all organisms studied thus far, the biosynthesis of all three aromatic amino acids begins with a common trunk of seven enzymes starting with the condensation of erythrose-4-P and phosphoenolpyruvate by DAHP synthase and ending with chorismate, the last intermediate common to all aromatic amino acids. From chorismate there are three major branches leading to L-phenylalanine, L-tyrosine, and L-tryptophan. This pathway possesses five enzymes that are commonly, but not always, targets of allosteric control. These enzymes are DAHP synthase, chorismate mutase (which converts chorismate to prephenate), and the enzymes catalyzing the initial irreversible step at the beginning of each of the three terminal branches.

Many means are used by microorganisms to ensure allosteric control of DAHP synthase (Jensen et al. 1967; Jensen and Nasser 1968; Jensen 1970; Jensen and Rebello 1970; Jensen and Stenmark 1970; Jensen and Twarog 1972). The discovery by Stenmark et al. (1974) of L-arogenate, then known as pretyrosine, in cyanobacteria was the initial hint that not only the regulation of but even the basic pathways to phenylalanine and tyrosine are highly variable in nature. Diversity (see Byng et al. 1982) of pathway construction and control has been documented by comprehensive studies of cyanobacteria (Hall et al. 1982) and of pseudomonads (Byng et al. 1980; Whitaker et al. 1981a, 1981b; Byng et al. 1983a, 1983b, 1983c). Five distinct rRNA homology groups of pseudomonads exist (Palleroni et al. 1973; Palleroni 1983), and we found that certain biochemical features of aromatic biosynthesis paralleled these five classes perfectly. Indeed, unknown species can be classified into these five groups and in some cases into distinct subgroups through the determination of patterns of enzyme arrangement and control (Byng et al. 1983b). Hence, once a
given pathway pattern is established to be conservative at some phylogenetic level, as gauged by rRNA, then the biochemical pattern can be used reliably for group placement.

The recent availability of dendrograms based on oligonucleotide catalogs has generated a new evolutionary thrust. We can now characterize biochemical pathways and their regulation in organisms that are known to have diverged from one another recently and then systematically work our way back to progressively earlier phylogenetic positions. Although our approach is to take the phylogenetic tree constructed by oligonucleotide cataloging at face value, the exact tree is undoubtedly subject to some revision. We fully expect that fine-tuning of branch placement will be in order for outlying dendrogram sections where phylogenetic distances are small. Such refinement can most conveniently and inexpensively be pursued by means of nucleic acid hybridization (Byng et al. 1983b). Indeed, if aromatic-pathway analysis yields patterns that lead to a questioning of the accuracy of certain branch points, additional information—including, perhaps, similar efforts with other metabolic systems—should help resolve these questions.

Current Interpretations about the Evolution of Aromatic Biosynthesis

A general outline of plausible evolutionary events that underlie aromatic-pathway biosynthesis and regulation is now feasible. Although only scattered information in a wide range of microbial groupings is available concerning aromatic-pathway biosynthesis and regulation, an extensive amount of background data exists concerning the division denoted as purple bacteria by Woese and co-workers (fig. 1).

Cofactor Specificity

Figure 1 depicts the simplest interpretation of the data thus far available about cofactor specificity of prephenate- and/or arogenate-dehydrogenase enzyme activities. In the frequent cases in which both dehydrogenase activities coexist, identical cofactor specificities have always been found. A strong rationale has been presented to argue that broad specificity for substrate (and cofactor) is a primitive property of enzymes (Jensen 1976). Such cofactor specificity for NAD+ or NADP+ is presented as the ancestral state that existed at the point of evolutionary divergence of the three superfamilies shown in figure 1.

So far, no instance of cofactor specialization has been found within superfamily A. In superfamily C, on the other hand, specialization for NADP+ occurred in an ancestor of the upper tripartite cluster (fig. 1) at an evolutionary time following divergence of that ancestor from group IV pseudomonads. In superfamily B, specialization for NADP+ occurred within the Acinetobacter lineage, in contrast to specialization for NAD+, which occurred in a common ancestor of Escherichia coli, group I, and group V pseudomonads. (Note the inappropriate contemporary naming of the two A. calcoaceticus strains within the same species; their oligonucleotide catalogs indicate a phylogenetic separation above the level of genus.)

Data concerning species within superfamilies A and C are too scattered to warrant interpretations of their aromatic-pathway evolution, but information about superfamily B is sufficiently comprehensive to suggest probable evolutionary progressions. Members of superfamily B thus far studied possess tyrosine-branch dehydrogenases that are specific for either NAD+ or NADP+. Do any superfamily
**COFACTOR SPECIFICITY OF AROMATIC PATHWAY DEHYDROGENASES**

![Dendrogram](image)

**Fig. 1.**—Dendrogram based on oligonucleotide cataloging of members of three superfamily components of the “purple bacteria”. The latter comprise one of eight major “eubacterial” lines of prokaryote descent depicted by Fox et al. (1980) within a lineage that is divergent from lineages containing archaeabacterial and eukaryotic kingdoms. Superfamilies A, B, and C correspond to purple nonsulfur-2, purple sulfur, and purple nonsulfur-1 groupings presented by Fox et al. (1980). The numbers positioned at points of divergence are $S_{AB}$ values obtained by C. R. Woese and co-workers. The dendrogram connections of the three superfamilies are drawn according to the dendrograms given by Palleroni (1983). Although the dendrogram given in figure 4 of Fox et al. (1980) shows superfamily A to connect at the deepest phylogenetic level ($S_{AB} = 0.30$), the deductions made about the evolution of cofactor specificity fit either tree equally well. *Rhizobium* strains were *Rh. leguminosarum* 3841 and *Rh. sp. KH486* from J. E. Beringer and A. Johnston. Other strains carry the following ATCC numbers: *Rhodopseudomonas capsulata*, ATCC 11166; *Rp. palustris*, ATCC 17001; *Rhodospirillum tenue*, ATCC 25093; *Rp. gelatinosa*, ATCC 17011; and *Rp. gelatinosa*, ATCC 17013. An extensive list of species comprising pseudomonad groups I, II, III, IV, and V are given in Byng et al. (1980), Whithaker et al. (1981a, 1981b), and Byng et al. (1983a). The $S_{AB}$ values given in this and succeeding figures were obtained from Fox et al. (1980), Stackebrandt and Woese (1981), Woese et al. (1982), and personal communication of unpublished data from G. Fox and C. R. Woese. The clusters of cofactor specificity thus far identified are shown by patterns defined in the box on the lower right.

B organisms retain the broad cofactor specificity consistent with the interpretation drawn in figure 1 for ancestral stem organisms of superfamily B? Because of its deep point of divergence within superfamily B (fig. 2), the *Desulfovibrio* lineage is a logical grouping for possible identification of retained breadth of cofactor specificity.
DAHP Synthase in Superfamily B

The dendrogram of superfamily B organisms, in relationship to the number of isoforms of DAHP synthase that are expressed in the lineages given, is shown in figure 3. An ancestral state of two isoforms is postulated to have already existed at this evolutionary time. Gene duplication leading to a third isozyme in the enteric lineage seems likely, whereas one of the two ancestral isoforms has been lost in group V pseudomonads (mainly species of Xanthomonas).

The DAHP synthase isoforms are named according to their specificities for feedback inhibitors, that is, DAHP synthase-tyr, DAHP synthase-trp, and DAHP synthase-phe. Group V pseudomonads possess only DAHP synthase-trp. Two-
EVOLUTION OF DAHP SYNTHASE ISOZYMES

**FIG. 3.—Dendrogram based on oligonucleotide cataloging of member groups within superfamily B in correlation with information available about 3-deoxy-D-arabino-heptulosonate 7-phosphate (DAHP) synthase. Solid branches indicate that two isoforms of DAHP synthase exist; barred branches, three isoforms: open branches, one enzyme. In *Escherichia coli* three differentially regulated isoforms of DAHP synthase are specified by cistrons *aroF*, *aroG*, and *aroH*, which map at 17 min, 37 min, and 57 min, respectively (Nachmann 1983). *Alteromonas putrefaciens* was fortuitously characterized in our pseudomonad studies (Whitaker et al. 1981a), having previously been misnamed as *Pseudomonas putrefaciens* *Aeromonas hydrophila* ATCC 9074, *A. liquefaciens* ATCC 14715, and *A. fumigatus* ATCC 13137 all have been shown to possess three DAHP synthase isoforms (Jensen, Nasser, and Nester 1967; Jensen and Stenmark 1970). Group I and group V pseudomonads were characterized in comprehensive detail for DAHP synthase (Whitaker et al. 1981a). Results shown for *Acinetobacter* strains are based on unpublished data. Nucleotide sequencing of *aroG* and *aroH* in *E. coli* K-12 have affirmed the origin of DAHP synthase-phe and DAHP synthase-trp from a common ancestral gene (Davies and Davidson 1982).

Isozyme organisms characteristically possess DAHP synthase-tyr and DAHP synthase-trp. DAHP synthase-phe occurs only in three-enzyme systems. Since DAHP synthase-phe seems to have been the most recently acquired isoform in superfamily B, it would be relatively easy to trace its origin within the enteric lineage. Is it present within the *Pasteurella/Photobacterium/Vibrio* lineage? If so, is it present within the little known but phylogenetically diverse *Oceanospirillum* lineage (fig. 2)?

Otherwise present throughout superfamily B, DAHP synthase-tyr became lost from the group V pseudomonad lineage. Since *Lysobacter* diverged from this group, the presence or absence of DAHP synthase-tyr in *Lysobacter* will pinpoint the evolutionary loss of this isoform before or after the point of divergence shown in figure 2.

DAHP synthase-trp has been detected in every superfamily B organism examined to date. A progression of DAHP synthase-trp enzyme types exists with respect to sensitivity to inhibition by chorismate. The enzyme in *E. coli* lacks sensitivity to inhibition by chorismate, in *Pseudomonas aeruginosa* it exhibits weak inhibition by chorismate, and in *X. campestris* it exhibits potent inhibition by chorismate (table 1). Feedback control of a single DAHP synthase by chorismate in *X. campestris* can be reconciled with a physiologically efficient pattern of regulation in which direct feedback inhibition of chorismate mutase and anthranilate synthase by end products would elevate chorismate levels, indirectly leading to inhibition of
Table I
Comparison of Allosteric Sensitivities of DAHP Synthase-trp from *Pseudomonas aeruginosa* with Those of the Single DAHP Synthase Enzyme of *Xanthomonas campestris*

<table>
<thead>
<tr>
<th>ORGANISM</th>
<th>PSEUDOMONAD GROUP</th>
<th>$K_i$ (L-TRYPTOPHAN) PEP</th>
<th>$K_i$ (CHORISMA) PEP</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Pseudomonas aeruginosa</em> <em>a</em></td>
<td>1</td>
<td>0.04 mM</td>
<td>1.35 mM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.005 mM</td>
<td>2.25 mM</td>
</tr>
<tr>
<td><em>Xanthomonas campestris</em> <em>b</em></td>
<td>5</td>
<td>0.38 mM</td>
<td>0.40 mM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.46 mM</td>
<td>0.10 mM</td>
</tr>
</tbody>
</table>

**NOTE.**—E-4-P = erythrose-4-phosphate; PEP = phosphoenolpyruvate.

* Data from Whitaker et al. (1982). Results are representative of other Group I organisms.

* Data from Whitaker et al. (1985). Results are representative of other Group V organisms.

DAHP synthase. This is a variation of a control pattern (Sequential Feedback Inhibition) that is well known in *Bacillus* species. Since in the latter case DAHP synthase and chorismate mutase activities coexist as a multifunctional protein (Byng and Jensen 1983), perhaps the allosteric binding sites for chorismate and L-tryptophan of the xanthomonad DAHP synthase arose by fusion of a gene for an unregulated DAHP synthase with the gene specifying an L-tryptophan-inhibited anthranilate synthase, after gene duplication of the anthranilate synthase cistron (Jensen 1976). If so, DAHP synthase-trp of both *P. aeruginosa* and *X. campestris* have already lost (author's unpublished data) intrinsic activity for anthranilate synthase (aminase); however, it is possible that data supporting such a gene-fusion event might yet be obtained from a study of *Acinetobacter* or *Desulfovibrio*.

Of interest is whether DAHP synthase-trp or DAHP synthase-tyr originated first. If DAHP synthase-trp was the ancestral enzyme, the physiologically important allosteric effector was probably chorismate. The subsequent evolutionary acquisition of DAHP synthase-tyr may have correlated with progressively increasing and decreasing sensitivities of DAHP synthase-trp to feedback inhibition by L-tryptophan and chorismate, respectively. If this scenario is correct, then DAHP synthase-trp of group V pseudomonads probably resembles the ancient evolutionary state, that is, that characterized by sensitive regulation by chorismate. On the other hand, if DAHP synthase-tyr originated first, the major sensitivity of DAHP synthase-trp to chorismate in group V pseudomonads may reflect an evolved regulatory adjustment to the loss of DAHP synthase-tyr. It should be informative to determine whether *Desulfovibrio* possesses one or two molecular species of DAHP synthase.

Metabolic Segment for L-Phenylalanine

The P-protein, having been found in every superfamily B organism studied to date, is clearly of ancient origin (fig. 4). This multifunctional protein carries both chorismate mutase and prephenate dehydratase activities and was first described in *Escherichia coli* and *Klebsiella pneumoniae* (Cotton and Gibson 1965). In *Acinetobacter calcoaceticus* prephenate dehydratase activity is heavily dependent on the presence of L-tyrosine (author's unpublished data). In group I pseudomonads prephenate dehydratase activity is stimulated ($K_m$ effect) by L-tyrosine (Byng et al. 1983c). In *E. coli* prephenate dehydratase activity is not activated by L-tyrosine at
all, even being inhibited by very high concentrations of L-tyrosine (Cotton and Gibson 1965). A comparison of the P-protein in *E. coli* with that in *P. aeruginosa* and *A. calcoaceticus* reveals similar enzymological characteristics. In *E. coli* the P-protein is active in the dimeric state, and the presence of feedback inhibitor (L-phenylalanine) promotes formation of an inactive tetramer (Baldwin et al. 1981). In *Acinetobacter* an active state requires L-tyrosine, and L-phenylalanine produces an inactive state of twofold greater molecular weight (author's unpublished data). It seems likely that these states correspond to dimer and tetramer moieties, respectively. *Pseudomonas aeruginosa* is intermediate in the sense that L-tyrosine activates prephenate dehydratase activity much less dramatically. Molecular weight transitions have not been studied in *P. aeruginosa* yet. An interesting continuum of diversity is thus emerging with respect to the effect of L-tyrosine on the prephenate dehydratase activity of the P-protein, and it should be revealing to characterize other P-proteins, selected with reference to dendrogram position, along these lines.

Figure 4 pictures an ancestral state in which the P-protein coexists with arogenate dehydratase and a monofunctional species of chorismate mutase F (Byng et al. 1983c). At an evolutionary time preceding an $S_{AB}$ value of 0.61 within the enteric lineage, arogenate dehydratase was lost. Although chorismate mutase F per se is absent from *E. coli* and *K. pneumoniae*, we presume that chorismate mutase F was the progenitor of the T-protein chorismate mutase prior
to the gene fusion with the gene encoding prephenate dehydrogenase that formed the multifunctional T-protein. Thus, *Alteromonas putrefaciens*, which predates evolution of the T-protein (fig. 5), still possesses chorismate mutase–F. *Acinetobacter* species, which lost arogenate dehydratase, also retained chorismate mutase–F as would be expected in the absence of an evolved T-protein.

Although wild-type *K. pneumoniae* ATCC 25304 lacks activity for arogenate dehydratase as does *E. coli*, a well-known derivative carrying multiple aromatic-pathway mutations has regained activity for arogenate dehydratase. *Klebsiella pneumoniae* 62-1, ATCC 25306 (Cotton and Gibson 1965) has been subjected to multiple mutagenesis protocols in order to obtain a triple auxotroph that has been widely used for accumulation of chorismic acid. If the active arogenate dehydratase of *K. pneumoniae* 62-1 is an atavistic enzyme resulting from reactivation of a cryptic gene that was initially silenced by a point mutation, then the enzyme would be expected to resemble the arogenate dehydratase of pseudomonad subgroup Ib species (Whitaker et al. 1981b). In conformity with this expectation, the *K. pneumoniae* enzyme will accept prephenate in addition to L-arogenate as substrate and is not feedback inhibited by L-phenylalanine. This arogenate dehydratase is synonymous with the “prephenate dehydratase A” noted by Cotton and Gibson (1965). Thus, wild-type *K. pneumoniae* appears to possess a pseudogene. Although only a few examples of such cryptic genes have been demonstrated (e.g., Lawther

**EVOLUTION OF L-TYROSINE BIOSYNTHE**

**Gene Duplication**

- **NAD^+**
  - SUPERFAMILY
  - 42
  - 40
  - **ATCC 23055**
  - **Acinetobacter calcoaceticus**
  - **ATCC 14987**
  - **Group V pseudomonads**
  - **Ib**
  - **Group I pseudomonads**
  - **77**
  - **Ia**
  - **Klebsiella pneumoniae**
  - **48**
  - **Escherichia coli**
  - **NADP^+**

**DIRECTION OF EVOLUTION**

**Symbol**
- Both prephenate and arogenate dehydrogenases are present.

**Explanation**

- **Fig. 5.**—Dendrogram based on oligonucleotide cataloging of member groups within superfamily B in correlation with information available about tyrosine biosynthesis and regulation. At an S_{AB} value of 0.39, two lineages separated, one having a dehydrogenase specific for NADP^+ and the other having a dehydrogenase that evolved specificity for NAD^+. The probability that duplication of the cistron encoding the primitive dehydrogenase occurred prior to the 0.42 bifurcation within the NAD^+ lineage is shown.
et al. 1981), their possible significance in evolution has been discussed recently (Hall et al. 1983).

One might predict that all enteric bacteria having an $S_{AB}$ relationship of 0.61 or higher with respect to $E. coli$ are likely to carry the cryptic gene for arogenate dehydratase. This could be tested by cloning the gene from an organism such as $P. aeruginosa$ (pseudomonad subgroup Ib) and using this as a molecular probe for detection of the cryptic gene. It also seems likely that restoration of an active arogenate dehydratase could be achieved by selective pressure using an appropriate mutant background in an organism such as $E. coli$.

Subgroup Ia pseudomonads lack both arogenate dehydratase and chorismate mutase--F (Byng et al. 1983c). The absence of chorismate mutase--F would appear to be a formidable barrier to the future evolution of a T-protein arrangement. Extensive genetic and physiological studies of $P. aeruginosa$, a subgroup Ib species possessing both of these activities, have established the roles of these activities in an unregulated, overflow pathway to L-phenylalanine (Fiske et al. 1983). The operation of this flow route seems to be intimately tied to the level of pathway flux permitted by any given carbon source.

Metabolic Segment for L-Tyrosine

Most members of superfamily B possess both prephenate dehydrogenase activity and arogenate dehydrogenase activity, as shown in figure 5. In the $Acinetobacter$ lineage a single enzyme of broad substrate specificity appears to account for both activities. Thus, (1) the two activities copurified during enzyme fractionation, (2) identical $K_m$ values for NADP$^+$ were calculated, and (3) identical $K_i$ values of 9 $\mu$M for L-tyrosine inhibition were measured.

The second lineage diverging at an $S_{AB}$ value of 0.39 exhibits dehydrogenase specialization for NAD$^+$ (see fig. 1). Since these dehydrogenases have been separated in group V pseudomonads (Whitaker et al. 1985) and in group I pseudomonads (Patel et al. 1978; author's unpublished data), it is likely that gene duplication of the cistron encoding an ancestral dehydrogenase of broad substrate specificity followed shortly after the event of cofactor specialization. Consistent with the presence of two enzyme species, $K_m$ values for NAD$^+$ differ depending on whether prephenate or L-arogenate is used as substrate, and $K_i$ values for L-tyrosine inhibition differ between prephenate dehydrogenase and arogenate dehydrogenase. In $X. campestris$ (Whitaker et al. 1985) one dehydrogenase is specific for prephenate, whereas the second dehydrogenase is reactive with either prephenate or arogenate. The latter dehydrogenase is presumably like the ancestral enzyme, whereas the former dehydrogenase has evolved specialized substrate specificity.

Within the enteric lineage, at a time following divergence from $A. putrefaciens$, arogenate dehydrogenase activity was lost. This evolutionary event was approximately coincident with the fusion of the cistrons encoding prephenate dehydrogenase and chorismate mutase--F, thereby creating the bifunctional T-protein. It is a striking contrast that the bifunctional T-protein of tyrosine biosynthesis is of recent origin within superfamily B, whereas the bifunctional P-protein of phenylalanine biosynthesis is very ancient. A systematic analysis of the organisms between $E. coli$ and $A. putrefaciens$ on the dendrogram of figure 2 should pinpoint the evolutionary origin of the T-protein. The foregoing scenario predicts that the presence of the T-protein will inevitably correlate with the absence of chorismate mutase--F.
Thus far, no aromatic-pathway dehydrogenase activity has been found to be insensitive to feedback inhibition by L-tyrosine. However, quantitative differences have been found to be consistent characteristics of different dendrogram sections; for example, pseudomonad group I (characterized by hypersensitivity of arogenate dehydrogenase to inhibition) can be reliably distinguished from pseudomonad group V (characterized by hypersensitivity of prephenate dehydrogenase to inhibition) on this criterion alone (Byng et al. 1980).

The Ancestral Superfamily B Pathway

The separate conclusions assembled in figures 3–5 concerning the evolutionary events involved in particular segments of aromatic amino acid biosynthesis can be combined to yield a view of the ancestral pathway of aromatic biosynthesis. The biosynthetic pathway that existed in the common ancestor of superfamily B organisms at an evolutionary time defined by an $S_{AB}$ value of 0.39 can be deduced as shown in figure 6. This ancestor possessed two isozymes of DAHP synthase, one sensitive to feedback inhibition by L-tyrosine and the other by L-tryptophan. The bifunctional P-protein, probably sensitive to L-phenylalanine-mediated feedback inhibition, had already evolved—being coexistent with an unregulated L-arogenate flow route to L-phenylalanine. Dual flow routes existed to L-tyrosine, but each of the two dehydrogenase reactions was catalyzed by a common enzyme. This broadly specific dehydrogenase also used either pyridine nucleotide cofactor and was probably already sensitive to feedback inhibition by L-tyrosine. No contemporary organism has yet been found that would fit the projected description of the superfamily B ancestor; it is most reminiscent of group I pseudomonads such as *Pseudomonas aeruginosa* (especially for phenylalanine synthesis) and of *Acinetobacter* species (especially for tyrosine synthesis).

Note in figure 6 that whereas gene duplication of the ancestral dehydratase (b) must have preceded acquisition of feedback inhibition, the evolution of feedback sensitivity in the ancestral dehydrogenase (c) apparently preceded gene duplication. Thus, all contemporary arogenate dehydratases so far studied within superfamily B (group I pseudomonads, group V pseudomonads, and *Klebsiella pneumoniae*) are unregulated (in contrast to prephenate dehydratases). On the other hand, both the activities of arogenate dehydrogenase and prephenate dehydrogenase within the NADP⁺ lineage of superfamily B have been found to be sensitive to feedback inhibition in every case evaluated thus far.

If the ancestral superfamily B pathway is regarded as an intermediate evolutionary state, a still earlier ancestral pathway can be estimated, as illustrated at the top of figure 6. The inference of this pathway is speculative, but when similar estimations become available for superfamilies A and C, a plausible ancestral pathway at the evolutionary stem of divergence of all three superfamilies will be deducible. The sole point of regulation shown in figure 6 (top) is at DAHP synthase. In organisms where allosteric regulation of aromatic biosynthesis is relatively simple (e.g., in cyanobacteria), early pathway control is commonly the sole or major regulation in force (Jensen and Hall 1982). If, as initial evolutionary events, L-tyrosine and L-tryptophan came to be utilized as end-product signals for early pathway control, then a selective pressure may have been created favoring preferential channeling of chorismate to L-phenylalanine. This would ensure that excess tyrosine
Duplication of cistrons coding [a] and [b]  
Fusion of new cistrons coding [a'] and [b'] = P-protein gene  
P-protein dehydratase becomes specific for PPA substrate  
P-protein dehydratase acquires sensitivity to feedback inhibition  
Dehydrogenase [c] acquires sensitivity to feedback inhibition

Fig. 6.—Aromatic biosynthesis and regulation deduced to exist in the ancestral superfamily B organism (bottom diagram) and (top diagram) an earlier hypothetical ancestor, showing the structures of chorismate (CHA), prephenate, (PPA), phenylpyruvate (PPY), 4-hydroxyphenylpyruvate (HPP), L-aroogenate (AGN), L-phenylalanine (PHE), L-tyrosine (TYR), and L-tryptophan (TRP). Enzyme denotations: a = chorismate mutase, b = dehydratase exhibiting broad specificity for prephenate or L-aroogenate, and c = dehydrogenase exhibiting broad specificity for prephenate or L-aroogenate and able to utilize either NAD+ or NADP+. The striped arrows indicate aminotransferase reactions that were catalyzed by a single enzyme. (Many contemporary aminotransferase enzymes are still broadly specific for the three substrates shown (Jensen and Hall 1982). Evolutionary steps intervening between the two evolutionary states are specified in the middle section between the two diagrams.
and tryptophan could not produce early pathway regulation that might provide starvation for phenylalanine. This may account for the ubiquity of the anciently evolved P-protein, in contrast to the rarity of the recently evolved T-protein.

The events during the evolutionary transition to the superfamily B ancestor that are pictured in figure 6—that is, duplication of the ancestral cistrons encoding chorismate mutase and the dehydratase, followed by gene fusion to account for the bifunctional enzyme—are those that would explain the evolution of the P-protein. The dehydratase component of the P-protein became specific for prephenate, and sensitivity to feedback inhibition evolved. The ancestral dehydrogenase, still broadly specific for either cofactor and either of the cyclohexadienyl substrates, probably evolved sensitivity to L-tyrosine-mediated feedback inhibition at this time.

Prospects for an Expanded Perspective on Aromatic-Pathway Evolution

The Major L-Tryptophan Branch

Obviously the tryptophan pathway must have coevolved in a very specific relationship with the phenylalanine and tyrosine pathways. Considerable biochemical and genetic diversity of tryptophan biosynthesis is indeed known to exist in nature (Crawford 1975). Of those organisms within superfamily B that have been studied thus far, extensive differences in tryptophan-pathway regulation and enzyme organization have been demonstrated, for example, in Escherichia coli vis-à-vis Pseudomonas aeruginosa (a group I b pseudomonad). Among the enteric organisms, E. coli and Klebsiella pneumoniae possess an organized complex of anthranilate synthase and phosphoribosyl transferase, whereas Serratia, Proteus, and Aeromonas species lack this enzyme complex. These results are in pleasing correspondence with the hierarchical evolutionary relationships of these organisms, as shown in figure 2. Inclusion of tryptophan-pathway studies in parallel with the study of phenylalanine and tyrosine biosynthesis would expand the base of information regarding a larger biochemical unit.

NAD+ Biosynthesis

A larger evolutionary insight with respect to aromatic biosynthesis will be possible through studies of vitamin-like compounds that originate from the aromatic pathway. Thus, Xanthomonas pruni (a group V pseudomonad) uses the aerobic tryptophan catabolic pathway for NAD+ biosynthesis, in contrast to E. coli, which uses the anaerobic dihydroxyacetone phosphate-aspartate pathway (Foster and Moat 1980). Hence, NAD+ biosynthesis is directly linked to aromatic biosynthesis in one superfamily B organism but not in another.

PABA Synthase

Different protein-protein arrangements of the two aromatic-pathway amidotransferases (anthranilate synthase and PABA synthase) are known to exist within superfamily B (see Byng et al. 1982). Both enzymes use an identical pair of substrates and have been hypothesized (Jensen 1976) to have originated from common ancestral genes encoding the aminase and glutamine-binding subunits. Recent results in the study of E. coli have indeed established homologous proteins: the nucleotide sequences of pabA and trp(G)D, which encode the aminase subunits, were 53% identical at the nucleotide level (Kaplan and Nichols 1983); the nucleotide sequences of pabB and trpE, which encode glutamine-binding subunits, were 40% homologous at the nucleotide level (Goncharoff and Nichols 1984).
Evolution of Biochemical Pathways

*Escherichia coli* exhibits a unique arrangement in which the glutamine-binding protein for anthranilate synthase (amidotransferase) is fused to the second tryptophan-pathway enzyme. In *P. aeruginosa* (a group I b pseudomonad), separate glutamine-binding subunits apparently exist for association with the aminase subunits of anthranilate synthase and PABA synthase. In *Acinetobacter calcoaceticus*, a single glutamine-binding subunit is shared by both amidotransferase activities. The latter system would conform most closely to an ancestral state in which a single pair of subunits catalyzed two reactions, with anthranilate likely being the major product and PABA being a minor (vitamin) product. A feasible evolutionary progression would be the following: (1) gene duplication of the ancestral aminase, (2) increased specialization of the two reactions, (3) gene duplication of the glutamine-binding subunit, and (4) gene fusion. Step (2) yields *A. calcoaceticus*; step (3) yields *P. aeruginosa*, and step (4) yields *E. coli*.

**Vitamin K**

Emerging data concerning the vitamin K pathway (Bentley and Meganathan 1982), another aromatic-pathway branch, suggests possibilities for still further biochemical diversity with respect to aromatic biosynthesis.

**Connecting Metabolic Networks**

Although biochemical pathways are commonly treated as separate entities, they are in fact part of a metabolic whole. The use of a pathway for aromatic amino acid biosynthesis may vary greatly in different organisms because of effects of secondary metabolism or catabolism in connecting pathways. If L-tryptophan is the precursor of great amounts of pigment (e.g., violacein in *Chromobacterium*), then the relative output of tryptophan will be much greater than in an organism like *Escherichia coli*. Microorganisms are highly diverse with respect to the use of aromatic amino acids for pigments, antibiotic peptides, and other secondary metabolites. It seems inescapable that a complete evolutionary appreciation of a given pathway must ultimately deal with connecting biochemical networks.

The distribution in nature of a particular metabolic ability is of interest. Until recently the important mammalian enzyme phenylalanine hydroxylase, which converts phenylalanine directly to tyrosine, was known only in *Pseudomonas* sp. ATCC 11299a (Guroff and Ito 1963), a strain which we (Berry et al. 1985) have shown to be *P. acidovorans*. This enzyme has recently been found in *P. facilis* (Friedrich and Schlegel 1972), *Alcaligenes eutrophus* (Friedrich and Schlegel 1972), and *C. violaceum* (Letendre et al. 1974), all of which are members of superfamily A. The full range of its distribution remains to be determined.

**Other Biochemical Pathways**

The branched pathway for biosynthesis of the aspartate family of amino acids is as complex in construction as the aromatic amino acid pathway and therefore exemplifies another system offering comparable potential for probing evolutionary relationships. Comparative studies done prior to the availability of phylogenetic trees (Cohen et al. 1969) indicate a degree of biochemical diversity approaching that found in the aromatic amino acid pathway. Entirely parallel evolutionary studies of two or more such systems in oligonucleotide-cataloged organisms would provide the beginnings of information that would allow interpretations of evolutionary
events that cross specific pathway boundaries. For example, the pyridine nucleotide cofactor domain of dehydrogenases is generally considered to be of ancient origin. Would the cofactor specificity for homoserine dehydrogenase in superfamily B organisms parallel that shown (fig. 1) for aromatic-pathway dehydrogenases? If so, what about dehydrogenase activities operating in other pathways?

New prospects for deducing evolutionary histories of biochemical pathways are by no means limited to branched pathways. Variation in any biochemical pathway is amenable to evolutionary conclusions when considered in the context of an acceptable phylogeny. An excellent example is the evolution of cell wall biosynthesis. A significant amount of biochemical data has been obtained to characterize many distinctive peptidoglycan types in microorganisms, and these results were obtained in oligonucleotide-cataloged strains. An example of the emerging evolutionary picture within one large microbial grouping is given in figure 3 of Stackebrandt and Woese (1981). Apparently, little information is yet available concerning the enzymological diversity that accounts for the observed diversity of peptidoglycan makeup.

Regulatory interactions are exerted between different biochemical pathways at the levels of both feedback inhibition (Jensen 1969) and transcriptional control (Bogosian and Somerville 1983). Such interlocking relationships have been difficult to study because they are subtle and not readily anticipated. Such relationships can now be probed in closely related microorganisms, since oligonucleotide-cataloging data provide a rational basis for strain selection. Hence, it should now be possible to study biochemical evolution at a level extending beyond the boundaries of a given pathway. Indeed, in the extent that the evolution of more and more particular pathways are deciphered in prokaryotes, the greater the prospects for the global evolutionary insights that will approach an ultimate understanding of biochemical evolution at the cell level.

Acknowledgments

I wish to acknowledge the intellectual input and benchwork contributions of R. Whitaker and A. Berry and to note especially the development of many ideas by G. S. Byng.

LITERATURE CITED


MATSATOSHI NEI, reviewing editor

Received April 27, 1984; revision received September 10, 1984.
A Comparative Description of Mitochondrial DNA Differentiation in Selected Avian and Other Vertebrate Genera

Louis G. Kessler and John C. Avise
Department of Genetics, University of Georgia

Levels of mitochondrial DNA (mtDNA) sequence divergence between species within each of several avian (Anas, Aythya, Dendroica, Melospiza, and Zonotrichia) and nonavian (Lepomis and Hyla) vertebrate genera were compared. An analysis of digestion profiles generated by 13–18 restriction endonucleases indicates little overlap in magnitude of mtDNA divergence for the avian versus nonavian taxa examined. In 55 interspecific comparisons among the avian congeners, the fraction of identical fragment lengths (F) ranged from 0.26 to 0.96 (F̄ = 0.46), and, given certain assumptions, these translate into estimates of nucleotide sequence divergence (p) ranging from 0.007 to 0.088; in 46 comparisons among the fish and amphibian congeners, F̄ values ranged from 0.00 to 0.36 (F̄ = 0.09), yielding estimates of p > 0.070. The small mtDNA distances among avian congeners are associated with protein-electrophoretic distances (D values) less than ~0.2, while the mtDNA distances among assayed fish and amphibian congeners are associated with D values usually >0.4. Since the conservative pattern of protein differentiation previously reported for many avian versus nonavian taxa now appears to be paralleled by a conservative pattern of mtDNA divergence, it seems increasingly likely that many avian species have shared more recent common ancestors than have their nonavian taxonomic counterparts. However, estimates of avian divergence times derived from mtDNA- and protein-calibrated clocks cannot readily be reconciled with some published dates based on limited fossil remains. If the earlier paleontological interpretations are valid, then protein and mtDNA evolution must be somewhat decelerated in birds. The empirical and conceptual issues raised by these findings are highly analogous to those in the long-standing debate about rates of molecular evolution and times of separation of ancestral hominids from African apes.

Introduction

Species and genera of birds commonly exhibit smaller genetic distances (D values) at protein-coding loci than do many nonavian vertebrates of same taxonomic rank (Martin and Selander 1975; Smith and Zimmerman 1976; Barrowclough and Corbin 1978; Avise et al. 1980a; Yang and Patton 1981; Zink 1982; Gutierrez et al. 1983). For example, from a multilocus electrophoretic survey of 10 species of waterfowl in the genus Anas, mean genetic distance was estimated to be 0.092

1. Key words: mitochondrial DNA, restriction endonucleases, evolutionary divergence, birds.
2. Present address: Department of Neurology, University of Texas Health Science Center, Dallas, Texas 75235.

Address for correspondence and reprints: John C. Avise, Department of Genetics, University of Georgia, Athens, Georgia 30602.

© 1985 by The University of Chicago. All rights reserved.
(range, 0.00-0.19; Patton and Avise 1985), while among 10 species of fish in the genus *Lepomis*, mean distance was 0.62 (range, 0.16-1.01; Avise and Smith 1977). Although ranges of genetic distance for various vertebrate congeners do overlap considerably, the general trend toward relatively small $D$ values in birds remains (reviewed in Avise and Aquadro [1982]). At higher taxonomic levels also, protein distances in birds (as assayed by immunological techniques) have proved to be unexpectedly small (Prager et al. 1974; Prager and Wilson 1980).

To account for this "conservative" pattern of protein differentiation in Aves, the following two fundamental alternative hypotheses (which occupy the end points on a continuum of possibilities) have been suggested (Zink 1982; Avise 1983): (1) relative to many nonavian vertebrates, avian taxa on the average have a more recent shared common ancestry and (2) the rate of protein evolution is decelerated in birds. The former hypothesis subsumes the possibility, forcefully argued by Sibley (1982), that birds are taxonomically "oversplit" at all supraspecific levels. Various empirical approaches may help decide between these competing hypotheses (Avise 1983). One approach involves evaluation of observed protein distances against available biogeographic or fossil evidence on absolute times of avian speciations. In two such case-history studies—involving waterfowl (Patton and Avise 1985) and North American warblers (Avise et al. 1980c)—it was tentatively concluded that avian protein evolution may indeed have been decelerated; however, the validity of this conclusion hinges critically on the reliability of divergence times, which in these two studies were taken from rather meager fossil and biogeographic evidence, respectively.

A second approach to helping distinguish between recent ancestry and protein deceleration involves examination of divergence in other portions of the avian genome. This is the approach employed in this study. If low levels of protein differentiation reflect recent common ancestry, a conservative pattern of differentiation should also characterize other aspects of the avian genome. Alternatively, if avian congeners are not especially young, the conservative pattern of substitution in replacement positions of protein-coding genes may be atypical of the remainder of the avian genome, only a small fraction of which has such protein-coding function (Shields 1983). For example, one theoretical possibility that might account for protein deceleration involves body temperature (Avise and Aquadro 1982). Conceivably, the relatively high and stable internal temperatures of most birds might provide a physiologic environment conducive to selection against certain amino acid substitutions (see Hochachka and Somero 1973). Such stabilizing selection need not extend to silent-position nucleotide substitutions or to regions of DNA not coding for proteins. While many other selectionist scenarios with varying theoretical predictions about genome differentiation might be entertained, it remains of primary importance to assess genome divergence empirically.

Here we employ restriction endonuclease fragment analyses to assess levels of mitochondrial DNA (mtDNA) differentiation in selected avian and other vertebrate genera. We have assayed representatives of five taxonomic groups for which background data on allozyme distances are also available: sunfish (*Lepomis*; Centrarchidae); waterfowl (*Anas, Aythya*; Anatidae); warblers (*Dendroica*; Emberizidae); sparrows (*Melospiza, Zonotrichia*; Emberizidae); and tree frogs (*Hyla*; Hylidae). Is the conservative pattern of differentiation in avian proteins also characteristic of the avian mitochondrial genome?
Background

Mitochondrial DNA in higher animals is a closed circular duplex molecule, maternally transmitted across animal generations. It is conserved in size (~15.7-19.5 kb) and gene content yet rapidly evolving in primary nucleotide sequence (Brown 1983). Direct sequencing and fine-scale mapping studies with a few organisms have shown that the majority of mtDNA evolution arises from base substitutions (primarily transitions) plus some very small addition-deletions. Large-scale additions, deletions, or rearrangements are uncommon (Aquadro and Greenberg 1983; Cann and Wilson 1983; Greenberg et al. 1983). In protein-coding regions of the molecule, silent-position changes greatly outnumber amino acid-replacing substitutions (Anderson et al. 1982; Brown and Simpson 1982; Cann et al. 1984). The rate of base substitution in mammalian mtDNA is reportedly at least five to 10 times higher than that of single-copy nuclear DNA (Brown et al. 1979, 1982).

Nucleotide sequencing is not yet practical for large-scale population surveys. However, restriction endonuclease–site mapping or fragment analyses can be used to estimate sequence divergence indirectly. In this study we employ a restriction-fragment approach to analyze mtDNA. In interpreting results of such analyses, two previously established relationships are of special relevance. (1) The first involves the mathematical relationship between \( F \) (the total proportion of shared fragments in the mtDNA digestion profiles of any compared samples) and \( p \) (the estimated nucleotide sequence divergence). The relationship between \( F \) and \( p \) is markedly curvilinear (Upholt 1977; Nei and Li 1979), such that for small values of \( F \) (i.e., <0.25), even small errors in estimating the fraction of identical fragment lengths (as might occur through chance electrophoretic comigration of nonhomologous fragments) will be reflected in large absolute differences in estimates of \( p \). (2) The second involves the empirical relationship between \( p \) and absolute divergence time \( t \), established on the basis of comparisons between 26 mammalian species pairs by Brown et al. (1979) and Brown (1983) (fig. 1). For \( p < -0.15 \) (corresponding to \( t = 8 \) Myr), mtDNA sequence divergence appeared linearly related to time, but for long divergence times \( p \) begins to plateau, until by 10 Myr “the readily-substituted positions in the mtDNA have become ‘saturated’ ” (Brown et al. 1982). Much of the remaining mtDNA is presumably under strong selective constraints (Aquadro et al. 1984). If the dynamics of mtDNA sequence differentiation generally proceed as indicated in figure 1, meaningful estimates of divergence can be attempted only when values of \( p \) are well within the linear portion of the curve, corresponding to \( F \) values > ~0.3.

Material and Methods

Between 13 and 18 informative restriction endonucleases were employed to assay the mtDNAs of various vertebrate species (table 1). Included in table 1 are only those enzymes found to produce multifragment digestion patterns in at least some species within each group. Enzymes that cleaved at no or one site in the mtDNAs of pairs of species being compared were excluded from the analyses. Primarily, five- and six-base-recognizing enzymes were used, because their mtDNA digestion profiles proved to be more readily interpretable than the more complex patterns produced by four-base enzymes.

Mitochondrial DNA was isolated from fresh tissue samples (heart, liver, or muscle) by the procedure of Lansman et al. (1981). In brief, the technique involves
homogenization of tissue, low-speed centrifugation to remove nuclei and debris, and subsequent lysis of mitochondria. MtDNA was purified by CsCl-ethidium bromide gradient centrifugation. Restriction endonuclease digestions of purified mtDNA were carried out under the vendor's (New England Biolabs) recommended conditions.

Digestion fragments were radioactively end-labeled using the large fragment of Escherichia coli DNA polymerase I and ³²P-αdCTP (Brown 1980) and then electrophoresed through agarose gels ranging in concentration from 0.6% to 2.2%. Digestion profiles were revealed by autoradiography of vacuum-dried gels (Maniatis et al. 1982). Fragment (and genome) sizes were determined by comparisons against molecular weight markers provided by HindIII digests of λDNA and PvuII-HincII double digests of pBR322.

For all waterfowl, warblers, sparrows, tree frogs, and some sunfish, digests were electrophoresed both at low and high gel concentrations to optimize resolution of large and small fragments, respectively. When necessary, electrophoresis was repeated to reorder samples for desired side-by-side comparisons. Fragment identity was assessed on the basis of comigration of fragments electrophoresed through the same gel. By including in our assays only those enzymes yielding digestion profiles with readily scorable numbers of fragments, by comparing fragments of questionable identity side-by-side on gels, and by employing a wide range of gel concentrations, we sought to minimize the possibility of judging as identical nonhomologous fragments that by chance comigrated.

The fraction of identical fragments was calculated for all pairwise comparisons of congeneric individuals by \( F = \frac{2N_{XY}}{N_X + N_Y} \), where \( N_X \) and \( N_Y \) are the numbers of fragments in genotypes \( X \) and \( Y \), and \( N_{XY} \) is the number of fragments shared. Values of \( F \) were converted to estimates of mtDNA nucleotide sequence divergence, \( p \), by the method of Nei and Li (1979), which involves weighting...
<table>
<thead>
<tr>
<th>Common Name (Scientific Name)* (Sample Size)</th>
<th>Restriction Endonucleases b</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Waterfowl:</strong></td>
<td></td>
</tr>
<tr>
<td>Mallard (Anas platyrhynchos) (2)</td>
<td></td>
</tr>
<tr>
<td>Green-winged teal (A. crecca) (3)</td>
<td></td>
</tr>
<tr>
<td>Mottled duck (A. fulvigula) (2)</td>
<td></td>
</tr>
<tr>
<td>Northern pintail (A. acuta) (4)</td>
<td></td>
</tr>
<tr>
<td>Gadwall (A. strepera) (2)</td>
<td></td>
</tr>
<tr>
<td>American wigeon (A. americana) (3)</td>
<td></td>
</tr>
<tr>
<td>Blue-winged teal (A. discors) (4)</td>
<td>1–9, 11–14, 16, 18</td>
</tr>
<tr>
<td>Northern shoveler (A. clypeata) (3)</td>
<td></td>
</tr>
<tr>
<td>Cinnamon teal (A. cyanoptera) (1)</td>
<td></td>
</tr>
<tr>
<td>Redhead (Aythya americana) (1)</td>
<td></td>
</tr>
<tr>
<td>Canvasback (A. valisineria) (2)</td>
<td></td>
</tr>
<tr>
<td>Lesser scaup (A. affinis) (3)</td>
<td></td>
</tr>
<tr>
<td>Ring-necked duck (A. collaris) (5)</td>
<td></td>
</tr>
<tr>
<td><strong>Sparrows:</strong></td>
<td></td>
</tr>
<tr>
<td>Song sparrow (Melospiza melodia) (4)</td>
<td></td>
</tr>
<tr>
<td>Swamp sparrow (M. georgiana) (3)</td>
<td></td>
</tr>
<tr>
<td>Lincoln’s sparrow (M. lincolni) (2)</td>
<td></td>
</tr>
<tr>
<td>White-throated sparrow (Zonotrichia albicollis) (4)</td>
<td>1–5, 7–19</td>
</tr>
<tr>
<td><strong>Warblers:</strong></td>
<td></td>
</tr>
<tr>
<td>Chestnut-sided warbler (Dendroica pensylvanica) (2)</td>
<td>1–5, 7–18</td>
</tr>
<tr>
<td>Blackburnian warbler (D. fusca) (1)</td>
<td></td>
</tr>
<tr>
<td>Magnolia warbler (D. magnolia) (2)</td>
<td></td>
</tr>
<tr>
<td>Cape May warbler (D. tigrina) (1)</td>
<td></td>
</tr>
<tr>
<td>Yellow-rumped warbler (D. coronata) (5)</td>
<td></td>
</tr>
<tr>
<td><strong>Sunfish:</strong></td>
<td></td>
</tr>
<tr>
<td>Redbreast sunfish (Lepomis auritus) (5)</td>
<td></td>
</tr>
<tr>
<td>Green sunfish (L. cyanellus) (7)</td>
<td></td>
</tr>
<tr>
<td>Pumpkinseed (L. gibbosus) (2)</td>
<td></td>
</tr>
<tr>
<td>Warmouth (L. galusus) (5)</td>
<td>1–6, 8–10, 12–14, 18</td>
</tr>
<tr>
<td>Dollar sunfish (L. marginatus) (2)</td>
<td></td>
</tr>
<tr>
<td>Longear sunfish (L. megalotis) (2)</td>
<td></td>
</tr>
<tr>
<td>Redear sunfish (L. microphalus) (4)</td>
<td></td>
</tr>
<tr>
<td>Spotted sunfish (L. punctatus) (4)</td>
<td></td>
</tr>
<tr>
<td>Bluegill (L. macrochirus) (10)</td>
<td></td>
</tr>
<tr>
<td><strong>Tree frogs:</strong></td>
<td></td>
</tr>
<tr>
<td>Bird-voiced tree frog (Hyla avivoca) (1)</td>
<td>1–14, 16, 18</td>
</tr>
<tr>
<td>Spring peeper (H. crucifer) (3)</td>
<td></td>
</tr>
<tr>
<td>Gray tree frog (H. chrysoscelis) (1)</td>
<td></td>
</tr>
<tr>
<td>Green tree frog (H. cinerea) (1)</td>
<td></td>
</tr>
<tr>
<td>Barking tree frog (H. gratiosa) (5)</td>
<td></td>
</tr>
</tbody>
</table>

* According to the most recent American Ornithologists' Union checklist (1982).

b Listed according to the following numerical designations (recognition sequences in parentheses): 1 = AvaI (CPyCGPuG), 2 = BsuIII (GGATCC), 3 = BstI (TGAAT), 4 = BseII (GCCN2GGO), 5 = BseII (AGATCT), 6 = BstII (GGTACC), 7 = ClaI (ATCGAT), 8 = HindII (GTPyPuAC), 9 = HindIII (AAGCTT), 10 = KpnI (GGTACC), 11 = NdeI (CATATG), 12 = PstI (CTGCAG), 13 = PvuII (CAGCTG), 14 = SacI (GAGCTC), 15 = SfiI (GGTACC), 16 = SsoI (AGGCTC), 17 = TfiI (TCGA), 18 = XbaI (TCTAGA), 19 = XmnI (GAAATTC).
Table 2
Estimates of mtDNA Differentiation among Four Species of Sparrows

<table>
<thead>
<tr>
<th></th>
<th>1 (Melospiza melodia)</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.026</td>
<td>0.030</td>
<td>0.068</td>
<td></td>
</tr>
<tr>
<td>2 (M. georgiana)</td>
<td>0.666</td>
<td>0.030</td>
<td>0.073</td>
<td></td>
</tr>
<tr>
<td>3 (M. lincolnii)</td>
<td>0.630</td>
<td>0.636</td>
<td>0.061</td>
<td></td>
</tr>
<tr>
<td>4 (Zonotrichia albicollis)</td>
<td>0.367</td>
<td>0.350</td>
<td>0.398</td>
<td></td>
</tr>
</tbody>
</table>

Note.—Results are based on restriction profiles from 18 endonucleases; \( F \) values among conspecifics were >0.93 (see text). Data above the upper-left to lower-right diagonal are nucleotide sequence divergence (\( p \)) values; those below the diagonal are total proportions of shared restriction fragments (\( F \)).

According to the numbers of fragments produced by four-, five-, and six-base enzymes, in addition, direct comparisons were made between the sparrow genera Melospiza and Zonotrichia and between selected representatives of the waterfowl genera Anas and Aythya (Kessler and Avise 1984).

Protein-electrophoretic distances (Nei’s \( D \) statistic [1972]) can be found in the following sources: waterfowl, Patton and Avise (1985); sparrows, Avise et al. (1980b); warblers, Avise et al. (1980c); sunfish, Avise and Smith (1977); treefrogs, Etges (1979). The mtDNA results have previously been presented (in the context of other evolutionary issues) for two groups: waterfowl (Kessler and Avise 1984) and sunfish (Avise and Saunders 1984).

Results

Matrices of mtDNA genetic differentiation for sparrows, warblers, and tree frogs are presented in tables 2, 3, and 4, respectively. Mean \( F \) values between species of Melospiza and between species of Dendroica were 0.65 (range, 0.63–0.67) and 0.52 (range, 0.44–0.61), respectively. These values are similar to previous estimates for the waterfowl genera Anas (\( \bar{F} = 0.41; \) range, 0.26–0.96) and Aythya (\( \bar{F} = 0.58; \) range, 0.51–0.65; Kessler and Avise [1984], tables 1 and 2) but are in marked contrast to the values found among Lepomis species of sunfish (\( \bar{F} = 0.10; \) range, 0.00–0.36; Avise and Saunders [1984], table 3) and species of Hyla tree frogs (\( \bar{F} = 0.07; \) range, 0.00–0.27; present study).

These differences are further reflected by the percentage of instances in which

Table 3
Estimates of mtDNA Differentiation among Five Species of Warblers

<table>
<thead>
<tr>
<th></th>
<th>1 (Dendroica pensylvanica)</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.043</td>
<td>0.035</td>
<td>0.031</td>
<td>0.048</td>
<td></td>
</tr>
<tr>
<td>2 (D. fusca)</td>
<td>0.529</td>
<td>0.044</td>
<td>0.055</td>
<td>0.052</td>
<td></td>
</tr>
<tr>
<td>3 (D. magnolia)</td>
<td>0.581</td>
<td>0.515</td>
<td>0.035</td>
<td>0.050</td>
<td></td>
</tr>
<tr>
<td>4 (D. tigrina)</td>
<td>0.612</td>
<td>0.436</td>
<td>0.585</td>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td>5 (D. coronata)</td>
<td>0.478</td>
<td>0.457</td>
<td>0.472</td>
<td>0.482</td>
<td></td>
</tr>
</tbody>
</table>

Note.—Results are based on restriction profiles from 17 endonucleases; \( F \) values among conspecifics were >0.92 (see text). Data above the upper left to lower right diagonal are nucleotide sequence divergence (\( p \)) values; those below the diagonal are total proportions of shared restriction fragments (\( F \)).
Table 4
Estimates of mtDNA Differentiation among Five Species of Tree Frogs

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Hyla avivoca</td>
<td>0.160</td>
<td>0.176</td>
<td>0.189</td>
<td>0.314</td>
<td></td>
</tr>
<tr>
<td>2 H. crucifer</td>
<td>0.098</td>
<td>0.253</td>
<td>0.600</td>
<td>0.237</td>
<td></td>
</tr>
<tr>
<td>3 H. chrysoscelis</td>
<td>0.268</td>
<td>0.056</td>
<td>0.189</td>
<td>0.316</td>
<td></td>
</tr>
<tr>
<td>4 H. cinerea</td>
<td>0.063</td>
<td>0</td>
<td>0.063</td>
<td>0.188</td>
<td></td>
</tr>
<tr>
<td>5 H. gratiosa</td>
<td>0.021</td>
<td>0.038</td>
<td>0.020</td>
<td>0.064</td>
<td></td>
</tr>
</tbody>
</table>

NOTE.—Results are based on restriction profiles from 16 endonucleases; F values among conspecifics were >0.83 (see text). In this table, some pairs of species exhibited F = 0 for six-base and/or five-base enzymes. For these species, estimates of sequence divergence are clearly large but unreliable in absolute magnitude, and the reported values (in parentheses) were generated under an arbitrary assumption that p = 0.6 for F = 0. Data above the upper-left-to-lower-right diagonal are nucleotide sequence divergence (p) values; those below the diagonal are total proportions of shared restriction fragments (F).

Congeneric species shared identical, multifragment digest profiles for particular enzymes. As shown in table 5, within the four avian genera, species sharing identical patterns ranged from 18% to 28% of all comparisons of digests; however, within Lepomis and Hyla, frequencies of profile sharing were only 0.8% and 2.4%, respectively. (Hyla crucifer was excluded from the comparisons because it differed obviously in genome size from the other Hyla species, as noted below.) In addition to exact sharing of many multifragment patterns, the avian groups commonly exhibited digestion profiles that could readily be interpreted as differing by the gain or loss of a single restriction site. This is exemplified by the autoradiographs in figures 2 and 3, which present HindIII digests of mtDNA from warblers and sparrows (see also fig. 1 of Kessler and Avise [1984] for HindIII patterns in waterfowl). In contrast, congeners within Lepomis (see Avise and Saunders 1984) and Hyla (NdeI digest, fig. 4) typically exhibit few fragment identities.

Only a limited attempt was made to estimate intraspecific differentiation of mtDNA in the groups surveyed. In pairwise comparisons of conspecific sparrows (numbers of individuals given in table 1), F values ranged from 0.93 to 1.0; among conspecific warblers, F values ranged from 0.92 to 1.0. These ranges of F values are similar to those observed among conspecific waterfowl (F = 0.93–1.0; Kessler and Avise [1984]). Since most conspecifics were collected at a single locale, our data almost certainly underestimate levels of mtDNA genetic variability within species. However, in two cases for which we do have limited geographic samples, surprisingly little differentiation among individuals was observed. Melospiza georgiana collected from Clarke County, Georgia, and San Petricio County, Texas, were indistinguishable as to mtDNA genotype by 16 restriction enzymes (F = 1.00). Two specimens of Dendroica coronata collected from these same two locales were also identical, while other individuals differed by a few mtDNA fragment changes (F = 0.92–1.0).

Our data on intraspecific variability within Hyla or Lepomis are also limited. However, it is noteworthy that among five H. gratiosa collected from a single locale, F values were as low as 0.83 and that no two individuals were identical in mtDNA genotype. Also, L. macrochirus is known to show extensive mtDNA sequence differentiation across the southern part of its geographic distribution, with F values between two subspecies equal to 0.32 (Avise et al. 1984). In the future it will be
Table 5
Summary of Various Measures of mtDNA Divergence in Bird, Sunfish, and Tree Frog Species

<table>
<thead>
<tr>
<th>TYPE OF COMPARISON</th>
<th>NUMBER OF PAIRWISE SPECIES COMPARISONS</th>
<th>NUMBER OF MULTIFRAGMENT DIGESTION COMPARISONS</th>
<th>PERCENTAGE SHARING OF MULTIFRAGMENT DIGESTION PATTERNS</th>
<th>MEAN NUMBER OF FRAGMENTS SCORED ( (N_x + N_y) ) PER SPECIES COMPARISON</th>
<th>MEAN (RANGE) OF ( F )</th>
<th>MEAN (RANGE) OF ( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birds:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Among congeneric species:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anas</td>
<td>36</td>
<td>483</td>
<td>18.2</td>
<td>108.4</td>
<td>0.41 (0.26–0.96)</td>
<td>0.062 (0.004–0.088)</td>
</tr>
<tr>
<td>Aythya</td>
<td>6</td>
<td>79</td>
<td>21.5</td>
<td>108.7</td>
<td>0.58 (0.51–0.65)</td>
<td>0.034 (0.025–0.043)</td>
</tr>
<tr>
<td>Melospiza</td>
<td>3</td>
<td>43</td>
<td>27.9</td>
<td>117.6</td>
<td>0.65 (0.63–0.67)</td>
<td>0.029 (0.026–0.030)</td>
</tr>
<tr>
<td>Dendroica</td>
<td>10</td>
<td>167</td>
<td>18.0</td>
<td>147.6</td>
<td>0.57 (0.44–0.61)</td>
<td>0.044 (0.031–0.055)</td>
</tr>
<tr>
<td>Between closely related genera:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anas/Aythya</td>
<td>2</td>
<td>27</td>
<td>3.6</td>
<td>115.0</td>
<td>0.19 (0.18–0.19)</td>
<td>0.109 (0.107–0.111)</td>
</tr>
<tr>
<td>Melospiza/Zonotrichia</td>
<td>3</td>
<td>45</td>
<td>2.2</td>
<td>129.0</td>
<td>0.37 (0.35–0.40)</td>
<td>0.067 (0.061–0.073)</td>
</tr>
<tr>
<td>Other vertebrates (among congeneric species):</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lepomis</td>
<td>36</td>
<td>378</td>
<td>0.8</td>
<td>70.4</td>
<td>0.10 (0.00–0.36)</td>
<td>0.277 (0.070–0.6)(^b)</td>
</tr>
<tr>
<td>Hyla</td>
<td>10</td>
<td>83(^a)</td>
<td>2.4</td>
<td>102.6</td>
<td>0.07 (0.00–0.27)</td>
<td>0.260 (0.156–0.6)(^b)</td>
</tr>
</tbody>
</table>

\(^a\) Excluding comparisons with \( H. \) crucifer.

\(^b\) Arbitrarily assumes \( p = 0.6 \) for \( F = 0 \) (see footnote to table 4).
Fig. 2.—HindIII digests of mtDNA isolated from sparrows. Lanes A–D depict results for Melospiza melodia; lanes E–G, M. georgiana; lanes H and I, M. lincolni; lane J, molecular-weight standards (sizes in kb indicated at left); and lanes K–M, Zonotrichia albicollis. Gel concentration is 1.5% agarose.

Fig. 3.—HindIII digests of mtDNA isolated from warblers. Lanes A and B depict results for Dendroica pennsylvanica; lane C, D. fusca; lanes D and E, D. magnolia; lane F, D. tigrina; lane G, molecular weight standards (sizes in kb indicated at left); and lanes H–L, D. coronata. Gel concentration is 1.5% agarose.
desirable to study more thoroughly the geographic differentiation in mtDNA in a variety of avian and other vertebrate species, particularly since Barrowclough (1983) reports that con specific birds generally exhibit little geographic heterogeneity in allozyme frequencies.

Table 5 summarizes various measures of mtDNA differentiation in the seven assayed vertebrate genera. Earlier (see Background), we discussed the assumptions that must be met in estimating mtDNA nucleotide sequence divergence, \( p \), from \( F \). For each of the avian groups, we have no evidence for large-scale additions or deletions of mtDNA sequence. Within the limits of resolution of our approach (±200 bp), mtDNA genome size appeared constant at \( \sim 16.5 \) kb for waterfowl and \( 17.3 \) kb for warblers and sparrows. The conservatism in mtDNA genome size for these groups is confirmed by the many enzymes for which the digestion profiles were either identical or readily interpretable by the apparent gain or loss of single sites. Such observations also support the assumption that identical fragments have usually been homologous.

In *Hyla* and *Lepomis*, attempts to compare genome sizes and accurately assess whether fragments of the same length were homologous were complicated (1) because most digestion patterns probably differed by multiple restriction site changes and (2) by the fact that very few multiband digestion profiles were shared across species. In one species we have unequivocal evidence for dramatic genome size variation. As judged by independent comparisons of many different digests (see, e.g., fig. 4), the mtDNA genome of *H. crucifer* is \( \sim 3 \) kb bases larger than those of

![Fig. 4.—*NdeI* digests of mtDNA isolated from tree frogs. Lane A depicts results for *Hyla avivora*; lane B, *H. crucifer*; lane C, *H. chrysoscelis*; lane D, *H. cinerea*; lane E, *H. gratiosa*; and lane F, molecular weight standards (sizes in kb indicated at right). Gel concentration is 0.7% agarose.](image-url)
the other assayed *Hyla*. We estimate its size to be \( \sim 23 \text{ kb} \), which would make it the largest mtDNA genome yet reported in a higher animal (Brown 1983).

Until direct sequence data become available, the estimates of \( p \) derived from fragment or site analyses must remain provisional. Nonetheless, in terms of fragment identificities in restriction-digest profiles, the assayed avian congeners clearly exhibit a conservative pattern of differentiation compared with the sunfish and tree frogs assayed. Even if it should prove true that the mtDNA digestion profiles of *Lepomis* or *Hyla* commonly differ because of large deletions, rearrangements, or other changes in addition to simple base substitution, this aspect of mtDNA differentiation itself would be of interest because of its contrast with the avian pattern.

**Discussion**

In recent years, several studies have begun to exploit the potential of the mtDNA genome for evolutionary analysis (reviewed in Avise and Lansman 1983; Brown 1983). Most projects have dealt with mammals, and the need for data on other vertebrates has been apparent (Brown 1983). The purposes of this study have been to: (1) expand the available data base on mtDNA differentiation, particularly for birds; (2) examine the empirical relationship between protein electrophoretic distance and mtDNA nucleotide sequence divergence, and (3) use these data to readdress the issues of genome conservatism and rate of molecular evolution in birds.

**Relative Magnitudes of mtDNA Sequence Divergence**

For the taxa considered in this study, mtDNA genetic distances among birds appeared smaller than those for the nonavian vertebrates (table 5 and fig. 5). For example, the largest mtDNA distance observed among congeneric birds in the genera *Anas*, *Aythya*, *Melospiza*, and *Dendroica* (a total of 55 pairwise species comparisons) was \( p = 0.088 \); only one distance value among *Lepomis* sunfish (36 interspecific comparisons) was less than this (\( p = 0.070 \)). In 10 comparisons among species of *Hyla*, the smallest distance estimate was \( p = 0.156 \). Furthermore, even distances between the avian genera *Anas* and *Aythya* (\( \bar{p} = 0.109 \)) and between *Melospiza* and *Zonotrichia* (\( \bar{p} = 0.067 \)) are lower than most distances between assayed sunfish or tree frogs within a single genus. We realize that the absolute mtDNA distances reported in this study may be subject to question, since nucleotide sequences were not determined directly and because several assumptions underlie the conversion of \( F \) to \( p \). Nonetheless, the relative ordering of mtDNA distances appears clear. Overall, in terms of relative magnitudes of mtDNA differentiation reflected in restriction-digest profiles, the avian groups appear to be “shifted down” approximately one taxonomic level compared to the nonavian groups studied so far.

This conclusion is further substantiated by another published study of mtDNA differentiation in birds. From cleavage map comparisons, Glaus et al. (1980) report \( p \) values ranging from 0.097 to 0.175 in comparisons **between genera** and subfamilies of galliform birds. These values are consistent with our estimate of \( \bar{p} = 0.109 \) between *Anas* and *Aythya* and fall within a range characteristic of many interspecific comparisons with *Lepomis*. They are also comparable to reported mtDNA distances (based on mapped sites) between two congeneric rat species, *Rattus rattus* and *R. norvegicus* (\( p = 0.137-0.184 \); Brown and Simpson [1981]), and between two
congeneric field mice, *Peromyscus maniculatus* and *P. leucopus* (*p* = 0.120–0.167; Avise et al. [1983]).

Empirical Relationship between *p* and *D*

The availability of estimates of both nuclear gene divergence (as measured by conventional protein-electrophoretic procedures) and mitochondrial sequence divergence for particular taxa permits comparisons between these distance measures. In figure 6 we have plotted *p* versus *D* for species pairs within *Anas*, *Aythya*, and *Lepomis*. For the avian congeners, all *p* values are <0.09, and these correspond empirically to *D* values in the range of 0.00–0.20. For *Lepomis*, most *p* values are >0.10, with associated *D* values ranging from 0.15 to 1.0 or more. Because of the nonlinear accumulation of sequence divergence in mtDNA with time beyond perhaps 8–10 Myr (see Background and fig. 1), Brown et al. (1982) suggest that study of relationships among organisms and estimates of absolute times of divergence be restricted to comparisons within the linear portion of the curve (where *p* < ~0.15–0.20). Data given in figure 6 suggest that such small *p* values appear to be associated with protein-electrophoretic distances of < ~0.20–0.40. Elsewhere we
have exploited these small mtDNA distances in waterfowl to assess systematic relationships within the group (Kessler and Avise 1984).

Much of the large variance in \( p \) values associated with \( D \) values >0.4 may be attributable to the mathematical relationship between scored fragment identity, \( F \), and \( p \) (see Background). In figure 6 we plot one example of a least-squares regression line relating \( p \) to \( D \). The dashed lines represent boundaries about this empirical regression assuming a ±10% error in estimation of \( F \) from which the \( p \) values were derived.

Divergence Times and Rates of Evolution in Birds

The conservative pattern of mtDNA differentiation in the avian versus nonavian taxa examined generally parallels the conservative pattern of protein differentiation reported previously. As argued in the Introduction, this result is compatible with the thesis that avian taxa may have a more recent shared common ancestry than do many comparable nonavian taxa. Can the magnitudes of mtDNA and protein divergence be reconciled with other information about absolute times of avian separation?

The fossil record for birds is notoriously poor, and it is usually not possible to accurately estimate separation times of particular pairs of extant species. However, previous interpretations of the somewhat better fossil remains for waterfowl indicated that both \( Anas \) and \( Aythya \) were already present by at least the Miocene epoch, that is, more than 15 Myr ago (Brodkorb 1964; Howard 1964; Romer 1966; Patton and Avise 1985). Available molecular data do not corroborate this interpretation even to a first approximation.
Brown et al. (1979, 1982) have calibrated a rate of nucleotide sequence divergence in mtDNA of \(\sim 2\% / \text{Myr} \). If we accept this rate, our mtDNA-estimated split between *Anas* and *Aythya* occurred \(\sim 5.5 \text{ Myr ago} \) \((\hat{d} = 0.109)\). Based on the allozyme data of Patton and Avise (1985), we have calculated a genetic distance of \(D = 0.164 \) (range, 0.113–0.313) between *Anas* and *Aythya* species. Using the slowest calibrated electrophoretic clock commonly employed in the literature (see Avise and Aquadro 1982), this genetic distance suggests a divergence time in the range of \(\sim 2.0–5.6 \text{ Myr ago} \). The protein- and mtDNA-based estimates are roughly comparable but are in sharp contrast to the fossil-based date of \(> 15 \text{ Myr} \).

If we accept the paleontological estimates of divergence time, the molecular data would argue for a decelerated rate of *both* mtDNA and protein evolution in waterfowl. However, avian paleontologists themselves have emphasized the provisional nature of many fossil assignments. “When evaluating the avian fossil record, it must be borne in mind that, in most instances, fossil species are known from only a few disarticulated bones. As the subfamilies and tribes of the Family *Anatidae* are not always clearly defined even in life, the allocation of extinct species to these groups, on the basis of one or two fragments of the skeleton, may quite properly be subject to question. . . . Many missing links must yet be found, and much more must be known of the osteology of living anseriforms before the fossil record can offer a true picture of the evolution of the group” (Howard 1964, pp. 235, 237).

Alternatively, if we question the reliability of the fossil assignments for waterfowl, a scenario arises that is analogous to the current debate over the divergence time of ancestral humans from African apes. Early interpretations of paleontological evidence suggested that humans and chimps last shared a common ancestor \(\sim 30 \text{ Myr ago} \) (Simons 1964, 1967; Pilbeam 1970). Subsequently, molecular data indicated that this split may have occurred as recently as 5 Myr ago (Sarich and Wilson 1967; Wilson et al. 1977; Andrews 1982; Andrews and Cronin 1982). While some evolutionists interpreted the molecular results to indicate a deceleration of sequence divergence in higher primates (Goodman 1976; Goodman et al. 1983), reevaluations of earlier paleontological studies and new fossil evidence have led many paleontologists to view more favorably the possibility of a recent human-chimp split (Johanson and White 1979; Greenfield 1980; Andrews 1982; Pilbeam 1984). Similarly, the notion that species of waterfowl (and of other avian genera) have speciated more recently than many nonavian vertebrate congeners should not be summarily dismissed. If the fossil evidence is indeed suspect, mtDNA- and protein-based estimates might be taken as more realistic indicators of avian divergence times.

Many possibilities remain. For example, it is conceivable that occasional introgression among hybridizing waterfowl or other avian species has inhibited genetic differentiation for some time after their “separation” (Prager and Wilson [1975] note that birds generally lose potential for interspecific hybridization slowly). Effects of such reticulate evolution in the early history of diverging species would be difficult to distinguish from more recent complete separation. It is also possible that the mtDNA rate calibrations, which were taken from mammalian data, will not apply to birds. As noted by Brown (1983), little is known about mtDNA replication repair in animals, and “rate of mtDNA evolution could vary considerably among various taxonomic groups.” If avian mtDNA is evolving at the mammalian rate, then protein evolution alone may be somewhat decelerated in birds (reconcil-
iation of mtDNA- and protein-based divergence times required use of a very slow protein clock). Overall, it is perhaps most likely that the final answer may include elements of both fundamental alternative hypotheses cited in the Introduction. Thus, rates of avian molecular evolution may be somewhat decelerated relative to other taxonomic groups, and many avian taxa are probably of more recent evolutionary age than their nonavian taxonomic counterparts. As with the human-chimp controversy, continued molecular and paleontological research will be required to shed additional light on these issues.

Acknowledgments

We thank Jonathan Arnold for advice with the data analysis. This work was supported by National Science Foundation grant BSR-8217291, by the Biomedical Research Support Grant Program of the National Institutes of Health, and by a National Institutes of Health Predoctoral Fellowship to L.G.K.

LITERATURE CITED


———. 1980c. Evolutionary genetics of birds. III. Comparative molecular evolution in New World warblers (Parulidae) and rodents (Cricetinae). J. Hered. 71:303–310.


SIBLEY, C. G. 1982. The relationships of the yellow-breasted chat (Icteria virens) and the alleged slowdown in the rate of macromolecular evolution in birds. Postilla 187:1–19.


MASATOSHI NEI, reviewing editor

Received July 20, 1984; revision received September 14, 1984.
'Tempo and Mode of Concerted Evolution in the L1 Repeat Family of Mice'

Sandra L. Martin, Charles F. Voliva, Stephen C. Hardies, Marshall H. Edgell, and Clyde A. Hutchison III
University of North Carolina

A 300-bp DNA sequence has been determined for 30 (10 from each of three species of mice) random isolates of a subset of the long interspersed repeat family L1. From these data we conclude that members of the L1 family are evolving in concert at the DNA sequence level in Mus domesticus, Mus caroli, and Mus platythrix. The mechanism responsible for this phenomenon may be either duplicative transposition, gene conversion, or a combination of the two. The amount of intraspecies divergence averages 4.4%, although between species base substitutions accumulate at the rate of ~0.85%/Myr to a maximum divergence of 9.1% between M. platythrix and both M. domesticus and M. caroli. Parsimony analysis reveals that the M. platythrix L1 family has evolved into a distinct clade in the 10–12 Myr since M. platythrix last shared a common ancestor with M. domesticus and M. caroli. The parsimony tree also provides a means to derive the average half-life of L1 sequences in the genome. The rates of gain and loss of individual copies of L1 were estimated to be approximately equal, such that approximately one-half of them turn over every 3.3 Myr.

Introduction

L1, or LINE1, is a large family of interspersed repeats found in mammals. Individual members of this family are bounded by short direct repeats on both ends and have a conserved 3' end which is characterized by an A-rich sequence (Voliva et al. 1984). The longest members of the repeat family are 6.5–7 kb in length; however, most copies are truncated at what appear to be random distances 5' to the A-rich region (Fanning 1983; Voliva et al. 1983, 1984). Implicit in this structure is the fact that the 3'-most sequences of the family are represented many more times in the genome than are the 5'-most sequences.

L1 has been characterized in greatest detail in primates (the KpnI family) and in rodents (the BamHI or MIF family) (see Singer [1982a, 1982b] for review). There are ~10^5 copies of the most abundantly represented sequences (Gebhard et al. 1982), which are interspersed throughout the genome. The function of the L1 repeats, like that of other interspersed repetitive families, is unknown. Sequences homologous to the L1 structure are found in RNA (Fanning 1982; Kole et al. 1983; Lerman et al. 1983; Shafit-Zagardo et al. 1983; Soriano et al. 1983), and evidence suggests that RNA polymerase II is responsible for their transcription (Shafit-
Recently, DNA sequence analysis has indicated that the L1 family may encode a polypeptide. The sequence of a long region that contains an open reading frame appears to be evolving under constraints imposed by selection for protein coding function (Martin et al. 1984).

Regardless of their function, restriction mapping and DNA hybridization experiments have indicated that members of the L1 family in rodents (Brown and Dover 1981) and in primates (Grimaldi and Singer 1983) are evolving in concert (Zimmer et al. 1980), i.e., copies of the repeat have greater similarity to one another when they are from the same species than they do when they are from different species. This behavior can be explained in two ways: (1) either the multiple copies are similar to one another within a species because they are the result of a recent amplification and dispersal of a progenitor family member or (2) the family was dispersed a long time ago and its members are constantly exchanging genetic information with one another by mechanisms such as unequal crossover or gene conversion.

This study was initiated with a goal of documenting whether or not members of the L1 family are evolving in concert at the DNA sequence level. The DNA sequence of random isolates of a 300-bp region of L1 was determined from three species of mice. This examination of how the sequences change in evolutionary time has provided new information about the relative rates of point substitution and genetic exchanges among members of the L1 repeat family. The rate of evolution by base substitution is \( \sim 0.85 \% / \text{Myr} \). Interestingly, the average half-life of individual copies in the genome is only 2–3 Myr, indicating that genetic exchange processes make an important contribution to the evolution of this multigene family.

**Material and Methods**

**Material**

Restriction enzymes, T4 polynucleotide kinase, and T4 polynucleotide ligase were purchased from New England Biolabs or Bethesda Research Labs. Calf intestinal phosphatase was prepared as described (Maniatis et al. 1982), and DNase I was purchased from Sigma. Enzymes were used in standard reaction conditions (Maniatis et al. 1982). Radioactive nucleotides were purchased from New England Nuclear. Liver from *Mus platythrix* was obtained from Litton Bionetics through Dr. Michael Potter.

**Methods**

Description of the construction and sequence determination of clones has been published elsewhere (Martin et al. 1984).

The single-stranded probe used to screen the pool of clones (an M13 vector with the Bam5 region from L1Md-2; Voliva et al. [1984]) was prepared by end-labeling phage DNA with \( \gamma^{32} \text{P}- \text{ATP} \). Five micrograms of the single-stranded DNA were treated with 2.5 ng/ml of DNase I in 6.6 M Tris–HCl (pH 7.5), 26.6 mM MgCl2, and 1.0 mM dithiothreitol for 10 min at 37°C, then extracted with phenol and precipitated with ethanol. The DNA was recovered and resuspended in 0.2 ml of 50 mM Tris–HCl (pH 8.1) and treated with 0.13 units of calf intestinal phosphatase for 1 h at 37°C. After phenol extraction and ethanol precipitation, the DNA was end-labeled with \( \gamma^{32} \text{P}- \text{ATP} \) using polynucleotide kinase (Maxam and Gilbert 1980), then separated from the unincorporated nucleotide by filtration through Sephadex G-50. Clones containing only one of the two possible orientations
of BamS in the M13 vector hybridize to this probe. Thus, all of the isolates have the same orientation as the insert and are ready for DNA sequencing without further characterization.

**Growth of the Recombinant Phage and Preparation of DNA for Sequencing**

An isolated plaque was inoculated into 0.5 ml of a fresh overnight culture of *Escherichia coli* strain JM103 (growth in minimal media; Miller [1972]). Thirty milliliters of YT media were added (Miller 1972), and the culture was grown at 37 C for 8 h. The bacteria were pelleted by centrifugation for 15 min in a Sorvall HS4 rotor, and 25 ml of the supernatant was recovered into a new tube containing 6.25 ml of 20% polyethylene glycol in 2.5 M NaCl. After sitting overnight on ice, the phages were pelleted by centrifuging for 20 min at 10,000 rpm in a Sorvall SS34 rotor. The pellet was carefully drained to remove excess polyethylene glycol, then resuspended in 2.25 ml of 10 mM Tris (pH 7.5), 1.0 mM EDTA, and 50 mM NaCl, and then extracted once with chloroform, once with phenol, and twice more with chloroform, or until the interface was clear. The aqueous phase was then precipitated with ethanol. The DNA was recovered in 0.2 ml of distilled water and precipitated again in 2.5 M ammonium acetate and ethanol, then resuspended in 50 μl of 1.0 mM Tris-HCl (pH 7.5) and 0.1 mM EDTA.

**Sequence analysis**

Differences between each pair of sequences were computed using the program seqdifM developed by C.A.H. and M.H.E. The maximum parsimony tree (Fitch 1977) was constructed with the help of a branch-swapping program (MPN; Czelusniak et al. [1982]) using a seed tree derived by the method of Fitch and Margoliash (1967). Direct examination of phylogenetically informative positions revealed few discordancies in the branching order because of the low frequency of base substitutions. Alternative seed trees consistent with each of the discordancies were used to ensure that a global minimum tree was found. Most of the ambiguities indicated in the tree are caused by a lack of substitutions in the interval between the origin of two sequences rather than by discordancies.

**Rate Calculation**

Experimentally, the average height of each node on the parsimony tree (fig. 5) was taken to be the age of one sampled sequence. The sampled ages were grouped into intervals corresponding to the time required to acquire four base substitutions. These intervals were determined by trial and error to be broad enough to absorb the variation in node heights that results from picking alternative trees from among the set of trees showing maximum parsimony. We analyzed the age distribution, making the assumption that the rates of gain and loss of repeats from the sampling pool are constant with respect to time. The constant rate of gain will mean that each age grouping was originally created with the same number of members as all of the others. The constant rate of loss will mean that each age grouping will have lost a constant proportion of its members as it passed through each successive time interval before arriving at its present age. These relationships will cause the age distribution to be a simple, exponentially declining, curve. The curvature will be a function of the rate of loss and is most simply characterized by a half-life.
Consideration of the limiting case in which the rate of loss is zero reveals that the intercept is the rate of gain.

The rates governing the age distribution of these repeats are a complex composite of the elementary rates for all processes affecting the gain and loss of repeats. Contributing factors include the speciation of *Mus domesticus* and *Mus caroli*, the rates of intragenomic exchange mechanisms that directly govern the repeats, and the laws of population genetics governing their fixation in the population.

**Results**

The Bam5 Region of Ll Is Found in Distantly Related Species of *Mus*

When genomic DNA from *Mus domesticus* is digested with the restriction endonuclease BamHI, size-fractionated by electrophoresis through agarose gels, and then stained with ethidium bromide, a band ~500 bp in length (Bam5) is visible over the background smear. This band corresponds to approximately 20,000 copies of a fragment contained within the larger repetitive structure of L1 (fig. 1). Hybridization of a Southern blot of BamHI-cleaved DNA from three species of *Mus*—*M. domesticus*, *M. caroli*, and *M. platythrix*—to a nick-translated Bam5 probe (originally isolated from the β-globin locus in the BALB/c mouse) demonstrates that all three species have this Bam5 repetitive unit (fig. 2). The intensity of hybridization to *M. platythrix* DNA, however, appears to be reduced. This suggests either that fewer of the repeats have retained the two BamHI sites that define the Bam5 fragment or that the similarity between the probe and *M. platythrix* DNA is reduced. Probably both of these conditions contribute, because the intensity of the 500-bp fragment is also reduced on staining with ethidium bromide.

**DNA Sequences Analysis of Random Isolates of Bam5**

Phage recombinants containing a single orientation of the Bam5 insert were isolated (see Material and Methods for experimental details). Three hundred nucleotides of DNA sequence were determined for each of 10 randomly isolated clones from each species. The sequence of the same region from two copies of the repeat, originally isolated from known locations within the β-globin locus of the BALB/c mouse (L1Md-2 and L1Md-4), was determined, in addition to these random isolates of Bam5. Figure 3 shows the nucleotide sequence of each isolate from all three species as well as a consensus sequence for each species.

The DNA sequences determined in this study are clearly homologous, differing at most by 12% in base substitutions between the two sequences p1 (from *M. platythrix*) and c6 (from *M. caroli*). The majority (68%) of the base substitutions were transitions. A few of the sequences had small deletions relative to the consensus sequence: in *M. domesticus*, isolate d7 had two 1-bp deletions; in *M. caroli*, c6 had a 17-bp deletion and c17 had a 1- and an 8-bp deletion. In *M. platythrix*, none of

![Fig. 1.—Structure of L1Md repeat. Sizes of the various restriction fragments are given in base pairs. Canonical positions for cleavage with EcoRI (R1) and BamHI (B) are shown. The 500-bp BamHI fragment near the 3' (right) end of the structure is the Bam5 fragment described in this study. The arrow indicates the direction and extent of sequencing. The 3' end is A-rich. The 5' end is not precisely defined for the longest members of the family.](image-url)
the 10 sequences had small deletions or insertions over the 300-bp region, although one additional isolate had an unusual organization. Positions 1–140 were unrelated to any sequence normally found in Bam5, but homology to Bam5 began at position 141, corresponding to position 210 in the other isolates. This isolate has not been included in the analysis of divergence. The simplest explanation of this structure is that it is one of the truncated copies of the repeat. Most (81%) of these random isolates contain an open reading frame throughout this 300-bp region (Martin et al. 1984).

Sequences from the Bam5 Region of L1 Are Evolving in Concert

The DNA sequence of each of the 32 isolates (30 random isolates plus two from the β-globin locus) was compared for 300 bp to every other isolate. The average divergence among all possible pairs was then taken for those comparisons involving sequences from the same species and those involving sequences from different species (table 1). Within the species *M. domesticus*, *M. caroli*, and *M. platythrix*, the average divergence between pairs of sequences is 4.2%, 4.9%, and 4.2%, respectively. (Gaps are not included as either mismatches, base substitutions, or positions available for comparison.) However, if pairs of sequences from different species are compared to one another, the average divergence between them is greater and correlates with the time of species divergence. Between the most closely related species, *M. domesticus* and *M. caroli*, the observed divergence, 5.4%, is only slightly greater than the divergence observed within each species. In contrast, comparison of both *M. domesticus* and *M. caroli* to their more distant relative, *M. platythrix*, reveals divergences of 9.1% and 9.2%, respectively. A tree summarizing these results

---

**Fig. 2.** Bam5 in genomic DNA from three species of *Mus*. Autoradiogram of a Southern blot with 5 μg of BamHI-cleaved DNA from *M. domesticus* (lane 1), *M. caroli* (lane 2), and *M. platythrix* (lane 3), after hybridization to nick-translated Bam5 DNA (originally isolated from L1Md-2; Voliva et al. [1984]). Positions of HindIII cleavage products of lambda DNA are marked and their sizes are given in kilobases. The O indicates the origin.
GGATCCGGAATACCTCTCCCTGGCCATATATCCAGAAGATGCCACCACGGGAAGAAGACAATGCTCCTATATGTTCATAGCAGCCCTTTTTATAATA dCON

GGATCCGGAATACCTCTCCCTGGCCATATATCCAGAAGATGCCACCACGGGAAGAAGACAATGCTCCTATATGTTCATAGCAGCCCTTTTTATAATA cCON

GGATCCGGAATACCTCTCCCTGGCCATATATCCAGAAGATGCCACCACGGGAAGAAGACAATGCTCCTATATGTTCATAGCAGCCCTTTTTATAATA pCON
Fig. 3.—DNA sequence of Bam5 isolates from *Mus domesticus* (top), *Mus caroli* (middle), and *Mus platyrrhix* (bottom). Sequences extend 300 bp, including the BamHI recognition sequence (GGATCC, beginning at position 1). The consensus sequence shows the most common nucleotide at each position for each species. Below the consensus sequence, positions that differ in each individual sequence are indicated: R = A or G, Y = C or T, and * = gap introduced to maximize the alignments. L1Md-2 and L1Md-4 are from the β-globin cluster of the BALB/c mouse (Voliva et al. 1983, 1984).
Table 1

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Number of Comparisons</th>
<th>Mean Divergence (%)</th>
<th>SD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within species:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Mus domesticus</em></td>
<td>66</td>
<td>4.2</td>
<td>2.3</td>
</tr>
<tr>
<td><em>M. caroli</em></td>
<td>45</td>
<td>4.9</td>
<td>2.1</td>
</tr>
<tr>
<td><em>M. platythrix</em></td>
<td>45</td>
<td>4.2</td>
<td>1.8</td>
</tr>
<tr>
<td>Between species:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>M. domesticus/M. caroli</em></td>
<td>120</td>
<td>5.4</td>
<td>1.3</td>
</tr>
<tr>
<td><em>M. caroli/M. platythrix</em></td>
<td>100</td>
<td>9.2</td>
<td>0.9</td>
</tr>
<tr>
<td><em>M. domesticus/M. platythrix</em></td>
<td>120</td>
<td>9.1</td>
<td>0.8</td>
</tr>
</tbody>
</table>

illustrates the known phylogenetic relationships of these three species of *Mus* and the divergence between each group of sequences (fig. 4). The divergence rate calculated from the divergence of *M. platythrix* from *M. domesticus* or *M. caroli* is 9.1% per $22 \times 10^6$ years or $4.1 \times 10^{-9}$ substitutions per site-year.

These data demonstrate that the Bam5 region of the L1 family is evolving in concert, since the sequences within each species are more closely related to each other than they are to the homologous sequences in different species. This effect is most striking in the comparisons involving sequences from *M. platythrix*. The divergence within that species is 4.2%. This contrasts with an average divergence of $\sim 9.1\%$ when *M. platythrix* sequences are compared to those of either *M. domesticus* or *M. caroli*.

![Figure 4](image-url)

**FIG. 4.**—Divergence of Bam5 within and between three species of *Mus*. Numbers are for the average of the observed percentage divergence between all possible pairs of sequences (see table 1), either within a species (below the species name) or between species (at the nodes). Interspecies divergence was converted to base substitutions per element by multiplying by 300/100 and then dividing by 2. The tree was then plotted according to the scale at left. The units on this scale correspond to branch lengths in figure 5. For comparison, a time scale (shown at right) is calibrated on the *M. platythrix/M. domesticus* plus *M. caroli* node. The approximate divergence times for these species are 10–12 Myr for *M. platythrix* to *M. caroli* and *M. domesticus* and 5–7 Myr for *M. domesticus* and *M. caroli* (V. Sarich, personal communication).
Phylogenetic analysis provides a more detailed picture of the relationships among these sequences. A parsimony analysis is able to resolve all of the Barn5 sequences from *M. platythrix* into a single clade that is distinct from the *M. caroli* and *M. domesticus* sequences (fig. 5). The sequences from *M. caroli* and *M. domesticus*, however, are not completely resolvable into distinct species-specific populations, indicating that some of the Barn5 sequences currently residing in these two species predate their divergence.

**Discussion**

We have demonstrated that sequences from the Barn5 region of the long interspersed repeat (L1) in mice are evolving in concert at the nucleotide sequence level. In the absence of some homogenizing process, the distances among repeats...
within and between species would have been the same. However, the observed
divergence between copies of the repeat is less within each species than it is between
them. Therefore, some type of genetic communication is occurring among the
individual copies of L1 within each species.

Mechanisms

Two mechanisms are consistent with the known features of the L1 family and
could account for its apparent concerted evolution: (1) the closely related copies of
L1 are the result of a recent dispersal of many new individuals from a common
source and (2) family members were dispersed long ago but are continually involved
in unequal crossovers or gene conversions that act to homogenize the population.

Recent amplification and dispersal, or duplicative transposition, of the sequences
of L1 analyzed in this study could explain the high degree of similarity observed
between family members among sequences from each species. An RNA and/or
cDNA intermediate may be involved in the dispersal of new copies (Jagadeeswaran
1984). There is good evidence for relatively recent insertions of L1 elements in mice
(Voliva et al. 1984) as well as in primates (Thayer and Singer 1983). The fact that
creatures as divergent as primates and rodents seem to have roughly equal numbers
of copies of L1 (Singer et al. 1983) suggests that it was also highly repetitive and
dispersed in their common ancestor. If recent amplification and insertion account
for their high degree of similarity, either the numbers of L1 individuals are increasing
all the time or there is a mechanism that removes copies as newer ones are inserted.
One might expect such a process to lead to noticeably different copy numbers
among different species, unless the rates of gain and loss are very closely balanced.

Unequal crossover and gene conversion mechanisms have been invoked to
explain the concerted evolution observed among members of a wide variety of
multigene families, including genes encoding ribosomal RNA (Arnheim et al.
1980] and references therein), histones, and globins (Slightom et al. 1980; Zimmer et al.
1980). Unequal crossover provides an appealing explanation for the homogenization
of repeat families whose members are found in tandem arrays. It does not, however,
account for homogenization among truly interspersed, highly repetitive structures
such as L1. Since unequal crossover relies on expansion, contraction, and sampling
to homogenize the sequences, such a mechanism cannot involve interspersed
sequences without the loss of intervening unique sequences. In addition, random
crossovers between copies that are scattered throughout the genome could rapidly
lead to a deleterious state of chromosomal translocations and deletions. A gene
conversion mechanism could skirt this difficulty, except that gene conversions and
unequal crossovers may be the result of different resolutions of the same initiating
event (Slightom et al. 1980). Thus, either there must be some way to suppress the
probability of a crossover or the above argument against crossovers is still an
important consideration. Another difficulty with the conversion model is that it
requires L1 sequences from literally all over the genome to pair and exchange
information in order to keep the entire dispersed family evolving in concert. Gene
conversion among these scattered copies of the repeat might be facilitated by the
use of a diffusible intermediate as the donor sequence. A candidate for such an
intermediate would be the RNA transcripts of L1 that have been detected, or
perhaps a cDNA copy of them.

The presence of short direct repeats flanking the copies of L1Md in the
β-globin gene cluster, as well as the recent insertion of L1Md-4 in the BALB/c
mouse (Voliva et al. 1984), indicate that duplicative transposition does occur in this repeat family. The intermediate in this process, presumably either a cDNA or an mRNA, could also act as a diffusible intermediate in a gene conversion process. These mechanisms and their relationship to concerted evolution are discussed in more detail in Dover (1982) and references therein. Given the present data, it is not possible to rule out either the dispersal mechanism or the conversion mechanism as processes used to maintain the similarity of the L1 family within a species. At this time it seems plausible that both mechanisms are involved. Resolution of this issue awaits analysis of truly orthologous copies of the repeat from genetically and evolutionarily well-characterized organisms.

Rates

Whatever the mechanism(s) used to maintain similarity between members of the L1 family within each species, the data obtained in this study provide information about the rate at which such genetic exchanges occur. *Mus platythrix* is thought to have diverged from *Mus caroli* and *Mus domesticus* ~10–12 Myr ago (Sarich, personal communication). In that period of time, their Bam5 sequences have become distinct from the related sequences in *M. caroli* and *M. domesticus* (fig. 5). This suggests that all the *M. platythrix* sequences have been replaced or inserted within the 10–12 Myr since *M. platythrix* last shared a common ancestor with *M. caroli* and *M. domesticus*. *Mus caroli* and *M. domesticus* are more closely related, diverging only ~5–7 Myr ago (Sarich, personal communication). In that period of time, most, but not all, of the Bam5 sequences have been homogenized. Thus, the time required to homogenize the population is somewhat greater than 5–7 Myr but less than 10–12 Myr. There are approximately 20,000 copies of the Bam5 fraction of the L1 repeat. If they were all replaced or inserted within the last 8 Myr, the minimum rate of that process must be >1.2 × 10^{-3} events per year. But this underestimates the true rate of genetic exchange because it assumes that all events are fixed and that each of the 20,000 copies is involved in only one event.

Using the branch lengths of individual sequences from the parsimony tree (fig. 5), a frequency distribution of repeats categorized by age was assembled (fig. 6). This distribution was used to derive estimates for the rates of gain and loss of the

![Graph](image)

**Fig. 6.—** Rate of turnover of L1 repeats during their concerted evolution. The frequency distribution shows the number of repeats sampled in each age interval corresponding to the time required to acquire four base substitutions. Each node in figure 5 was taken to be the birth of a new repeat, and the age of that repeat was taken to be the average node height in substitutions. The number of nodes falling in each interval is given. For comparison, a curve is shown for a population of repeats with a constant total number and in which one-half of the elements are replaced by new ones during each time interval.
Concerted Evolution in the L1 Repeat Family

repeats during their concerted evolution (see Methods subsection). The rate of loss conforms to a half-life corresponding to four base substitutions. For comparison, a curve was drawn exactly representing this half-life. The rate of gain is also such that approximately half of the repeats appear in a similar time interval. This is true because half of the observed repeats fall in the first time interval. The rate of gain is approximately equivalent to the rate of loss, in accordance with a steady-state situation in which the total copy number of repeats does not change dramatically with time. The rates were found to be such that one-half of the repeats are lost and replaced by new copies in the time it takes to acquire approximately four base substitutions. Using our divergence rate of $4.1 \times 10^{-9}$ substitutions per site-year and a size of 300 sites per element, we calculate that this corresponds to a turnover of half of the repeats every 3.3 Myr. Thus, the genetic exchange processes that underlie the turnover of L1 repeats in the genome are playing a major role in the evolution of this large multigene family. Because these repeats are found in such high copy numbers and are interspersed throughout the genome, their high rate of flux may provide important avenues for affecting plasticity in the organization and expression of the genome in evolutionary time.

Acknowledgments

Sandra L. Martin is a fellow of The Jane Coffin Childs Memorial Fund for Medical Research. Charles F. Voliva was supported by predoctoral training grant GM07092 from the National Institutes of Health to the University of North Carolina. This research was supported by Public Health Service grants AI08998 and GM21313 from the National Institutes of Health.

LITERATURE CITED


GRIMALDI, G., and M. F. SINGER. 1983. Members of the KpnI family of long interspersed
repeated sequences join and interrupt α-satellite in the monkey genome. Nucleic Acids Res. 11:321–338.


Coevolution of the Glucose Dehydrogenase Gene and the Ejaculatory Duct in the Genus *Drosophila*

Douglas R. Cavener

Department of Molecular Biology, Vanderbilt University

The glucose dehydrogenase gene (Gld) in *Drosophila melanogaster* exhibits a unique spatial and temporal pattern of expression. Gld expression switches from a non-sex-limited state at the pupal stage to a male-limited state at the adult stage. At the adult stage, the enzyme is restricted to the ejaculatory duct. Within the genus *Drosophila*, the ejaculatory duct has undergone a simple morphological divergence. In order to determine whether correlated changes in Gld expression had occurred, Gld activity during the pupal and adult stages was determined for several *Drosophila* species. It was found that virtually all of the species exhibit pupal Gld activity, whereas only those species with an expanded ejaculatory duct express male-limited Gld. The results of interspecific genital imaginal disc transplantation experiments indicate that the expanded morphology and Gld expression do not require any species- or sex-specific diffusible factors. An apparent regulatory polymorphism exists within the *D. takahashi* species with respect to male-limited Gld expression.

Introduction

During the past decade there has been a growing awareness that changes in gene regulation may be the major vehicle of organismal evolution (Wilson 1976; Raff and Kaufman 1983). Considerable progress has recently been made in understanding the evolution of alcohol dehydrogenase regulation in the genus *Drosophila* (Rabinow and Dickinson 1981; Batterman et al. 1983; Benyajati et al. 1983). These studies indicate that gene and promoter duplication and the divergence of cis-acting regulatory elements have occurred within the genus *Drosophila* to effect a differential pattern of ADH expression. Such studies have discovered a multiplicity of solutions for the same basic tissue and developmental pattern of expression in different species. In contrast, few studies have focused on the coordinate change in morphology and gene regulation in a closely related taxonomic group. Unfortunately, few systems have been identified that are amenable to study.

I have identified a gene in *Drosophila* that is ideally suited for investigating the coevolution of morphology and gene regulation. This gene encodes the structural gene for glucose dehydrogenase (Gld; EC 1.1.99.10; Cavener [1980]). (The name and acronym for the glucose dehydrogenase structural gene were mistakenly identified earlier as glucose oxidase and Go, respectively; Cavener and MacIntyre [1983]). In *Drosophila melanogaster*, Gld is expressed during the pupal stage where it is required for eclosion (i.e., emergence from the puparium after the completion of metamorphosis). The expression of Gld becomes male limited at the adult stage.

Key words: Coevolution, *Drosophila*, glucose dehydrogenase gene, male reproductive organ, morphology.
stage, whereupon it is transferred to females during copulation (Cavener and MacIntyre 1983). Of particular importance to this report is that GLD is expressed only in the ejaculatory duct of males during the adult stage. We have found that pupal and male GLD are encoded by a single gene. Thus, the D. melanogaster Gld gene exhibits an unusually restricted pattern of temporal and spatial regulation. In order to gain an understanding of the evolution of GLD regulation, I have examined the developmental expression of GLD and the morphology of the ejaculatory ducts of several Drosophila species representing the major species groups.

Material and Methods

With the exception of Drosophila melanogaster and simulans, all other species listed in tables 1–3 were obtained from the National Drosophila Species Resource Center, Bowling Green University, Bowling Green, Ohio. The D. melanogaster strain used in these studies was the MWI (Cavener and MacIntyre 1983) wild-type strain. The D. simulans strain was isolated from a natural population in Georgia in 1979. All species were reared on a standard cornmeal medium at 22–24 C. Pupae were isolated from the walls of half-pint bottles and classified into the following stages by microscopic examination: prepupae (PP); stage 1 pupae (P1, no eye pigment visible); stage 2 pupae (P2, light eye pigment visible); stage 3 pupae (P3, dark eye pigment); stage 4 pupae (P4, wings and bristles pigmented); and stage 5 pupae (P5, full pigmentation of pharate adult). Male and female adults were isolated from stock cultures. The sexes were separated and placed in fresh vials seeded with live yeast. The adults were maintained at 22–24 C for 3–7 days before their GLD activity was assayed.

Spectrophotometric GLD assays were performed after methods described by Cavener and MacIntyre (1983) except that 0.05 M potassium phosphate buffer, pH 6.8, was used for both the homogenization and assay buffer. These assays were performed with a Gilford 2600 UV/VIS spectrophotometer with a 7225 Hewlett-Packard X-Y Plotter. Each data point represents the mean of three replicates from a homogenate of 6–12 individuals except for data in table 4, where the body part of only a single individual was used.

The genital imaginal disc transplantation experiments were performed after the procedures of Ursprung (1967). Genital imaginal discs from male melanogaster third instar larvae were isolated using tungsten dissecting needles. Single discs were injected into either male or female pseudoobscura third instar larvae using an injection needle forged from 20-μl Drummond microcaps. Four to 5 days after eclosion, the pseudoobscura hosts were dissected. The genital imaginal disc derivatives from the melanogaster male donor were assayed for GLD.

Results and Discussion

GLD activity in Drosophila melanogaster is restricted to third instar larvae, prepupae, pupae, and adult males (Cavener and MacIntyre 1983). In order to determine whether the preadult GLD developmental profile exhibited in melanogaster was generally found in the genus Drosophila, three other Drosophila species were examined in detail. The three species chosen (ananassae, pseudoobscura, and immigrans) represent a varying degree of evolutionary relatedness to melanogaster. Ananassae, pseudoobscura, and melanogaster are in the Sophophora subgenus (Throckmorton 1962). However, ananassae and melanogaster are more closely
related. *Immigrans* is within an entirely different subgenus (*Drosophila*). The preadult GLD activity profile is quite similar for all four species (fig. 1). GLD activity increases from third instar larval stage and reaches peak activity during the last three pupal stages. These data are consistent with the hypothesis that the function of *melanogaster* GLD in eclosion (Cavener and MacIntyre 1983) is conserved in the *Drosophila* genus.

A more extensive phylogenetic survey of the preadult GLD activity was done using only stages P3–P4 pupae (table 1). Most species exhibit substantial levels of pupal GLD activity. Because many species pupate in the medium, it was impractical to determine the wet weight of the pupae. Thus, the pupal GLD activity is given only in units per individual. Unfortunately, interspecific comparisons of pupal GLD activities are not meaningful because of the variation in size among these species. It is obvious from inspection that pupal size and adult body size are highly correlated within species. Thus, meaningful intraspecific comparison between GLD pupae and adult activities can be made when activity is computed in units per individual. Table 1 lists the pupal, adult male, and adult female GLD activities in units per individual. In all but three species (*kikkawai, willistoni, and saltans*), pupal activity is more than twofold higher than adult female GLD activity. This comparison is important because it generally appears that the low apparent GLD activity of females is largely an artifact of our assay. (See below for further discussion on this point.) Thus, the female activities may approximate the background noise of our assay.

Because *Drosophila* adult females are generally larger than adult males, the GLD activities of adult males and females are given in units of wet weight for comparison (table 2). The most important feature of these data is derived from the comparison of male and female adult GLD activities. Female GLD activities (table 2) are uniformly low for the 26 species assayed. In contrast, two rather distinct male GLD activity groups were found. One group (seven species) exhibits high male

![Fig. 1.—Glucose dehydrogenase developmental profile of *Drosophila melanogaster* (○), *D. pseudoobscura* (■), *D. immigrans* (□), and *D. ananassae* (▲) from third instar larvae to the termination of metamorphosis. Each data point represents the mean of three to six replicates. 3L = late third instar larvae. PP = prepupae. P1–P5 = stage 1–5 pupae as defined in Material and Methods.](image-url)
Table 1
Glucose Dehydrogenase Activity in Various *Drosophila* Species

<table>
<thead>
<tr>
<th>Group and Species</th>
<th>Pupae</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Melanogaster</em> melanogaster</td>
<td>93.8 (2.55)</td>
<td>201.8 (3.07)</td>
<td>8.9 (0.82)</td>
</tr>
<tr>
<td><em>Melanogaster</em> simulans</td>
<td>41.6 (0.75)</td>
<td>158.1 (4.43)</td>
<td>5.5 (2.00)</td>
</tr>
<tr>
<td><em>Melanogaster</em> yakuba</td>
<td>20.7 (0.95)</td>
<td>21.3 (1.28)</td>
<td>4.8 (1.28)</td>
</tr>
<tr>
<td><em>Melanogaster</em> mauritiana</td>
<td>61.7 (1.64)</td>
<td>78.5 (2.18)</td>
<td>14.2 (0.59)</td>
</tr>
<tr>
<td><em>Melanogaster</em> rjasekari</td>
<td>43.3 (0.50)</td>
<td>24.1 (0.63)</td>
<td>10.6 (0.28)</td>
</tr>
<tr>
<td><em>Melanogaster</em> eugracilis</td>
<td>33.2 (1.01)</td>
<td>601.8 (3.55)</td>
<td>11.8 (1.32)</td>
</tr>
<tr>
<td><em>Melanogaster</em> pseudotakahashii</td>
<td>30.6 (1.18)</td>
<td>22.4 (0.46)</td>
<td>7.2 (0.87)</td>
</tr>
<tr>
<td><em>Melanogaster</em> kikkawai</td>
<td>19.2 (0.16)</td>
<td>7.6 (0.49)</td>
<td>11.6 (0.56)</td>
</tr>
<tr>
<td><em>Melanogaster</em> birchii</td>
<td>20.5 (0.63)</td>
<td>3.8 (0.28)</td>
<td>9.5 (2.68)</td>
</tr>
<tr>
<td><em>Melanogaster</em> elegans</td>
<td>20.6 (0.39)</td>
<td>6.1 (0.74)</td>
<td>4.4 (0.80)</td>
</tr>
<tr>
<td><em>Melanogaster</em> ficusphila</td>
<td>50.4 (2.19)</td>
<td>17.3 (3.26)</td>
<td>22.4 (2.93)</td>
</tr>
<tr>
<td><em>Melanogaster</em> ananassae</td>
<td>54.9 (5.74)</td>
<td>10.3 (0.71)</td>
<td>12.4 (1.28)</td>
</tr>
<tr>
<td><em>Obscura</em> pseudoobscura</td>
<td>115.4 (1.07)</td>
<td>9.2 (0.37)</td>
<td>9.4 (0.61)</td>
</tr>
<tr>
<td><em>Willistoni</em> willistoni</td>
<td>11.4 (0.51)</td>
<td>14.5 (1.08)</td>
<td>13.3 (0.63)</td>
</tr>
<tr>
<td><em>Saltans</em> saltans</td>
<td>14.9 (0.63)</td>
<td>3.9 (0.57)</td>
<td>14.1 (0.42)</td>
</tr>
<tr>
<td><em>Immigrans</em> immigrans</td>
<td>46.8 (0.76)</td>
<td>13.6 (0.80)</td>
<td>59.6 (7.92)</td>
</tr>
<tr>
<td><em>Tripuncata</em> tripuncata</td>
<td>68.2 (0.76)</td>
<td>1.9 (0.57)</td>
<td>5.3 (0.28)</td>
</tr>
<tr>
<td><em>Guarani</em> subbadia</td>
<td>97.9 (1.26)</td>
<td>8.8 (0.72)</td>
<td>12.4 (0.67)</td>
</tr>
<tr>
<td><em>Guarani</em> quarani</td>
<td>49.7 (5.20)</td>
<td>13.6 (1.24)</td>
<td>18.0 (0.68)</td>
</tr>
<tr>
<td><em>Cardini</em> cardini</td>
<td>35.5 (0.70)</td>
<td>10.4 (3.20)</td>
<td>11.8 (0.76)</td>
</tr>
<tr>
<td><em>Repleta</em> mojavensis</td>
<td>74.5 (0.80)</td>
<td>10.2 (0.44)</td>
<td>8.3 (0.25)</td>
</tr>
<tr>
<td><em>Repleta</em> mercatorum</td>
<td>145.9 (2.64)</td>
<td>15.7 (1.36)</td>
<td>26.0 (2.05)</td>
</tr>
<tr>
<td><em>Repleta</em> repleta</td>
<td>169.0 (2.50)</td>
<td>28.8 (1.24)</td>
<td>43.6 (2.81)</td>
</tr>
<tr>
<td><em>Repleta</em> hydei</td>
<td>97.4 (1.37)</td>
<td>33.4 (3.02)</td>
<td>41.6 (1.38)</td>
</tr>
<tr>
<td><em>Virilis</em> virilis</td>
<td>79.8 (4.28)</td>
<td>10.8 (2.24)</td>
<td>10.0 (0.45)</td>
</tr>
<tr>
<td><em>Robusta</em> robusta</td>
<td>108.1 (2.11)</td>
<td>13.5 (2.83)</td>
<td>10.9 (0.63)</td>
</tr>
</tbody>
</table>

*NOTE—SDs of the activity measurements are given in parentheses; μmol DCIP reduced min⁻¹ fly⁻¹ ml⁻¹ × 10⁻⁵.*

GLD activity. The ratios of male to female GLD activity range from 3.7 to 71.4 in this group. Interestingly, all seven species are in the *melanogaster* species group. Four of these species (*melanogaster*, *simulans*, *yakuba*, and *mauritiana*) are within the *melanogaster* species subgroup as well. The second group of species exhibit low GLD activity, and their male/female GLD activity ratios are close to 1. All male/female ratios are less than 2 except for *D. robusta* (ratio = 2.3). This group contains species from virtually all of the major *Drosophila* species groups including the *melanogaster* group (*kikkawai*, *birchii*, *elegans*, and *ficusphila*).

The low apparent GLD activity exhibited by all of the females and the males of the second group are of questionable importance. Because the GLD assays are performed with crude homogenates, it is difficult to know whether the low activities are the result of catalysis of GLD or of a glucose-dependent reduction of DCIP by other unidentified enzymes. Adult females from a *melanogaster* Gld mutant strain that totally lack the *Gld* gene have almost as much apparent "GLD activity" as wild-type *melanogaster* females (unpublished data). However, very faint electromorphs can be detected from tenfold concentrates of *robusta* and *willistoni* adults. These electromorphs are largely derived from nonreproductive tissues. These data taken together suggest that a very low level of GLD activity may be present in all adults of all *Drosophila* species. Despite the popular notion that most genes may be
Table 2
Glucose Dehydrogenase Activities in Monophasic and Biphasic Species

<table>
<thead>
<tr>
<th>Species</th>
<th>Male</th>
<th>Female</th>
<th>Male/Female</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High male activity group:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>melanogaster*</td>
<td>258.7</td>
<td>8.0</td>
<td>32.3</td>
</tr>
<tr>
<td>simulans*</td>
<td>247.0</td>
<td>5.8</td>
<td>42.6</td>
</tr>
<tr>
<td>yakuba*</td>
<td>48.4</td>
<td>7.6</td>
<td>6.4</td>
</tr>
<tr>
<td>mauritiana*</td>
<td>126.6</td>
<td>15.4</td>
<td>8.2</td>
</tr>
<tr>
<td>rajasekaran*</td>
<td>34.9</td>
<td>9.4</td>
<td>3.7</td>
</tr>
<tr>
<td>eugracilis*</td>
<td>771.5</td>
<td>10.8</td>
<td>71.4</td>
</tr>
<tr>
<td>pseudotakahashi*</td>
<td>32.0</td>
<td>6.7</td>
<td>4.8</td>
</tr>
<tr>
<td><strong>Low male activity group:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kikkawai</td>
<td>15.2</td>
<td>13.0</td>
<td>1.2</td>
</tr>
<tr>
<td>birchii</td>
<td>8.4</td>
<td>17.9</td>
<td>0.5</td>
</tr>
<tr>
<td>elegans</td>
<td>11.1</td>
<td>7.5</td>
<td>1.5</td>
</tr>
<tr>
<td>ficusphila</td>
<td>23.1</td>
<td>20.4</td>
<td>1.1</td>
</tr>
<tr>
<td>ananassae</td>
<td>13.2</td>
<td>15.1</td>
<td>0.9</td>
</tr>
<tr>
<td>pseudoobscura</td>
<td>10.0</td>
<td>7.5</td>
<td>1.3</td>
</tr>
<tr>
<td>willistoni</td>
<td>24.2</td>
<td>14.3</td>
<td>1.7</td>
</tr>
<tr>
<td>salvans</td>
<td>8.9</td>
<td>13.1</td>
<td>0.7</td>
</tr>
<tr>
<td>immigrans</td>
<td>6.0</td>
<td>17.9</td>
<td>0.3</td>
</tr>
<tr>
<td>tripuncata</td>
<td>2.1</td>
<td>3.7</td>
<td>0.6</td>
</tr>
<tr>
<td>subbadia</td>
<td>5.2</td>
<td>5.8</td>
<td>0.9</td>
</tr>
<tr>
<td>guarani</td>
<td>9.5</td>
<td>9.7</td>
<td>1.0</td>
</tr>
<tr>
<td>cardini</td>
<td>10.3</td>
<td>8.3</td>
<td>1.2</td>
</tr>
<tr>
<td>mojavensis</td>
<td>11.3</td>
<td>6.6</td>
<td>1.7</td>
</tr>
<tr>
<td>mercatorum</td>
<td>14.4</td>
<td>17.2</td>
<td>0.8</td>
</tr>
<tr>
<td>repleta</td>
<td>15.7</td>
<td>20.4</td>
<td>0.8</td>
</tr>
<tr>
<td>hydei</td>
<td>17.7</td>
<td>15.0</td>
<td>1.2</td>
</tr>
<tr>
<td>virilis</td>
<td>5.8</td>
<td>4.3</td>
<td>1.3</td>
</tr>
<tr>
<td>robusta</td>
<td>6.2</td>
<td>2.1</td>
<td>2.5</td>
</tr>
</tbody>
</table>

**NOTE.**—μmol DCIP reduced min⁻¹ mg wet weight⁻¹ ml⁻¹ × 10⁻³.

* Melanogaster subgroup.

b Suzukii subgroup.

e Eugracilis subgroup.

t Takahashi subgroup.

entirely transcriptionally inactive during certain developmental periods or in certain tissues, there are no known methods to rigorously test this notion. Furthermore, there are ample theoretical reasons for believing that genes have a finite probability of being transcribed even under conditions when such genes are thought to be in a general state of repression.

Despite the difficulty in the interpretation of very low activity values, the high adult male GLD activity values exhibited by melanogaster, simulans, mauritiana, and eugracilis are nonetheless impressive. Examination of the ejaculatory ducts of the various species studied yielded an important clue to their high activities. All of the high-GLD-activity males have an enlarged ejaculatory duct. The expanded state of the ejaculatory duct is relatively rare among the genus Drosophila. Throckmorton (1962) originally documented this morphological difference (fig. 2). Among the low-activity species, only guarani males exhibit an expanded ejaculatory duct. The expansion of the guarani ejaculatory duct is quite small and occurs only at the
Fig. 2.—Pictoral phylogeny of paragonia, vasa deferentia, and the anterior ejaculatory duct in Drosophila. 1) cancellata, 2) lacicola, 3) repleta, 4) species O, 5) guarani, 6) mediostrata, 7) quinaria, 8) subfunetris, 9) immigrans, 10) neocordata, 11) equinoxialis, 12) simulans, 13) ananassae, 14) algonquin, 15) populi. Copied with permission of Dr. Lynn Throckmorton. The anterior ejaculatory duct in each drawing is the unpaired structure attached laterally to the paragonia (large paired structures) and vasa deferentia (smaller paired structures). Note the expanded state of the melanogaster group ejaculatory duct.

extreme anterior end. Indeed, my observations indicate that the guarani ejaculatory duct is considerably less expanded than indicated in Throckmorton's drawings (fig. 2). The importance of the correlation between male GLD activity and the morphology of the ejaculatory duct is dependent on the assumption that GLD is synthesized in the ejaculatory duct. Besides the fact that GLD activity is localized in the ejaculatory ducts of melanogaster males, the results of the imaginal disc transplantation experiments (described below) are consistent with the hypothesis that GLD is synthesized in the ejaculatory duct.

Coincidentally, the first takahashii strain that I obtained from the Drosophila Species Resource Center proved to be exceptional; it had an expanded ejaculatory duct but lacked GLD activity. Six other geographical strains of takahashii were obtained and examined (table 3). All seven strains have an expanded ejaculatory duct. The variation in pupal activity (table 3) is relatively small. Interestingly, male activities among the seven strains show a continuum. Two strains, both from the Philippines, have male/female GLD ratios close to 1. Two other strains (Nepal and South India) have ratios greater than 3. Distinct GLD electromorphs have been visualized for the Nepal, South India, and the two Taiwan strains. I have not been able to detect distinct GLD electromorphs from the two strains from the Philippines. The Philippine strains likely contain cis- and/or trans-acting regulatory mutations that result in the lack of GLD expression in the ejaculatory duct. A mutation in the coding region of Gld is unlikely since these strains have normal pupal GLD.
Table 3
Glucose Dehydrogenase Activity in Drosophila takahashii Strains

<table>
<thead>
<tr>
<th>Strain</th>
<th>Pupae*</th>
<th>Male</th>
<th>Female</th>
<th>Male/Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philippines, 3146.13</td>
<td>33.8 (0.84)</td>
<td>9.4 (0.31)</td>
<td>8.2 (0.75)</td>
<td>1.2</td>
</tr>
<tr>
<td>Philippines, 3146.14</td>
<td>24.3 (0.80)</td>
<td>11.4 (0.56)</td>
<td>10.2 (1.38)</td>
<td>1.1</td>
</tr>
<tr>
<td>Nepal, 2363.4</td>
<td>37.7 (1.22)</td>
<td>33.7 (0.33)</td>
<td>8.7 (0.82)</td>
<td>3.9</td>
</tr>
<tr>
<td>South India, 3072.1</td>
<td>26.9 (1.48)</td>
<td>25.6 (0.51)</td>
<td>7.4 (0.93)</td>
<td>3.5</td>
</tr>
<tr>
<td>Taiwan, 3146.16</td>
<td>35.5 (0.39)</td>
<td>35.1 (0.18)</td>
<td>12.3 (0.76)</td>
<td>2.9</td>
</tr>
<tr>
<td>Thailand, 3252.7</td>
<td>44.7 (0.72)</td>
<td>11.4 (0.07)</td>
<td>5.8 (0.71)</td>
<td>2.0</td>
</tr>
</tbody>
</table>

NOTE.—μmol of DCIP reduced min⁻¹ mg of wet wt⁻¹ ml⁻¹ × 10⁻⁵.
* Pupal activities are given in units per individual rather than mg of weight.

activities. However, this hypothesis assumes that a single Gld gene exists in takahashii, as is the case for melanogaster (unpublished data).

The function of pupal GLD is most likely conserved given its required function for eclosion in melanogaster (Cavener and MacIntyre 1983). However, the apparent pupal GLD activities of a few species (e.g., willistoni and saltans) are so low that it is questionable whether GLD is functional at the pupal stage in these species.

The limited number of species expressing significant male GLD activities and possessing an expanded ejaculatory duct suggests that these two characters coevolved to serve an ancillary fertility function. Specifically, I propose that the ejaculatory duct, in most Drosophila species, functions solely as a propulsive organ, whereas in a few species it has an expanded size and an additional role to include a secretory function. According to this proposal, GLD is one of the secretory products of the ejaculatory duct in such species. It is uncertain whether GLD activity is an absolute requirement for male fertility in melanogaster. Gld null mutants that are thought to be due to point mutations are male fertile. However, a Gld null mutant that totally lacks the Gld gene (by virtue of two overlapping deletions) is male sterile. These contradictory results may arise from the presence of a very small amount of GLD activity that supports wild-type function in the putative point mutants, or it may be that the sterility of the Gld deletion mutant is caused by the mutation of another gene. At present I cannot rule out the null hypothesis that the expanded ejaculatory duct and GLD serve no functions in adult males.

One possible hypothesis for the observed correlation between the ejaculatory duct morphology and male GLD activity is simply that the quantity of activity is directly related to the size of the ejaculatory duct. Two facts weigh against this hypothesis. (1) Although the ejaculatory ducts in the high-activity group are larger than those in the low-activity group, the size differential is relatively small (fig. 2) when compared with the large differences in GLD activity (table 2). (2) Visual inspection of the size of the ejaculatory ducts within the high-activity group clearly indicates that there is not a positive correlation between size and GLD activity. For example, the takahashii strains have exceptionally large ejaculatory ducts but have rather low levels of male GLD activity compared with other species with smaller ejaculatory ducts (e.g., melanogaster and simulans). These data are more consistent with the hypothesis that the ejaculatory duct morphology and the male-limited expression of GLD coevolved. We have recently isolated genomic DNA clones of...
the *melanogaster* and *pseudoobscura* *Gld* genes. We intend to use these clones and *Gld* clones from other species to test this hypothesis by comparing DNA sequences and interspecific DNA-mediated gene transformation experiments.

There are several possible explanations for the high male *GLD* activity exhibited by a few species in the *melanogaster* species group. The major explanations include (1) unique cis- and/or trans-acting regulatory elements and (2) unique hormonal and/or developmental environments. I have used the method of imaginal disc transplantation to determine if the specific hormonal or developmental milieu of *melanogaster* males are requirements of male *GLD* expression. The male genital imaginal disc gives rise to all of the internal and external reproductive structures except for the testes. Among these structures is the ejaculatory duct. Imaginal discs, transplanted to a host third instar larvae, can undergo complete differentiation in parallel with the corresponding host disc. This technique has provided a means of testing the cellular autonomy of development and gene expression in *melanogaster* (Ursprung and Nothiger 1972).

Although a few reports of successful interspecific transplantation experiments exist (Nothiger 1964) for closely related *Drosophila* species, there was no assurance that transplantation between *pseudoobscura* and *melanogaster* would be successful. Several reciprocal transplantation experiments were attempted between these two species. To date, the transplantation experiments involving the injection of male *pseudoobscura* male discs into *melanogaster* hosts have been unsuccessful. I have not recovered differentiated structures derived from the *pseudoobscura* donor. The reciprocal experiment has been successful in two cases. One experiment involved a female *pseudoobscura* host and the other involved a male *pseudoobscura* host. In both cases, the donor male *melanogaster* disc differentiated fully into normal male genital structures. Of critical importance, the *melanogaster* ejaculatory duct exhibited its normal expanded state. Thus, the unique morphology of the *melanogaster* ejaculatory duct is under cell autonomous control of the genital imaginal disc. *GLD* activity analysis of these ejaculatory ducts indicates that they have activities that are characteristic of the *melanogaster* species (table 4). It should be noted that ejaculatory ducts lose much of their *GLD* activity during the isolation processes. However, *melanogaster* ejaculatory ducts reared in *pseudoobscura* hosts (table 4, C and D) have similar levels of *GLD* activity to the *melanogaster* controls (table 4, A and B) and substantially more activity than the *pseudoobscura* host tissues (e.g., head and thorax). Similar results have been obtained from experiments involving male

| Table 4 |
| Glucose Dehydrogenase Activity in Various Organs and Body Parts |

<table>
<thead>
<tr>
<th>Host</th>
<th>Tissue</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. melanogaster, male</td>
<td>melanogaster, ejaculatory duct</td>
<td>24.8</td>
</tr>
<tr>
<td>B. melanogaster, male</td>
<td>melanogaster, ejaculatory duct</td>
<td>18.0</td>
</tr>
<tr>
<td>C. pseudoobscura, female</td>
<td>melanogaster, ejaculatory duct</td>
<td>24.3</td>
</tr>
<tr>
<td></td>
<td>pseudoobscura, head and thorax</td>
<td>4.7</td>
</tr>
<tr>
<td>D. pseudoobscura, male</td>
<td>melanogaster, ejaculatory duct</td>
<td>13.3</td>
</tr>
<tr>
<td></td>
<td>pseudoobscura, head and thorax</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>pseudoobscura, reproductive organs</td>
<td>0</td>
</tr>
</tbody>
</table>

*NOTE.*—μmol of DCIP reduced min⁻¹ tissue⁻¹ ml⁻¹ × 10⁻⁵.
melanogaster donors and female melanogaster hosts (unpublished data). These experiments also indicate that GLD is synthesized by one of the organs derived from the imaginal disc, most likely the ejaculatory duct itself.

In summary, these experiments demonstrate that the morphology of the ejaculatory duct and the expression of GLD in this organ does not require any species- or sex-specific hormonal or developmental factors that come from sources other than organs derived from the genital imaginal disc. It is quite likely, then, that the regulatory and developmental differences observed are largely the result of cis- and trans-acting factors that exist in the ejaculatory duct itself.

LITERATURE CITED


RICHARD K. KOEHN, reviewing editor

Received June 19, 1984; revision received September 24, 1984.
A New Method for Estimating Synonymous and Nonsynonymous Rates of Nucleotide Substitution Considering the Relative Likelihood of Nucleotide and Codon Changes

Wen-Hsiung Li, Chung-I Wu, and Chi-Cheng Luo

University of Texas Health Science Center at Houston

A new method is proposed for estimating the number of synonymous and nonsynonymous nucleotide substitutions between homologous genes. In this method, a nucleotide site is classified as nondegenerate, twofold degenerate, or fourfold degenerate, depending on how often nucleotide substitutions will result in amino acid replacement; nucleotide changes are classified as either transitional or transversional, and changes between codons are assumed to occur with different probabilities, which are determined by their relative frequencies among more than 3,000 changes in mammalian genes. The method is applied to a large number of mammalian genes. The rate of nonsynonymous substitution is extremely variable among genes; it ranges from $0.004 \times 10^{-9}$ (histone H4) to $2.80 \times 10^{-9}$ (interferon $\gamma$), with a mean of $0.88 \times 10^{-9}$ substitutions per nonsynonymous site per year. The rate of synonymous substitution is also variable among genes; the highest rate is three to four times higher than the lowest one, with a mean of $4.7 \times 10^{-9}$ substitutions per synonymous site per year. The rate of nucleotide substitution is lowest at nondegenerate sites (the average being $0.94 \times 10^{-9}$), intermediate at twofold degenerate sites ($2.26 \times 10^{-9}$), and highest at fourfold degenerate sites ($4.2 \times 10^{-9}$). The implication of our results for the mechanisms of DNA evolution and that of the relative likelihood of codon interchanges in parsimonious phylogenetic reconstruction are discussed.

Introduction

Until the late 1970s the most frequently used method for estimating the numbers of nucleotide substitutions between two homologous DNA sequences had been the one by Jukes and Cantor (1969; see also Holmquist 1972; Kimura and Ohta 1972). These authors made the simplifying assumption that nucleotide substitution occurs randomly. In recent years the rapid accumulation of DNA sequence data and the realization of nonrandom patterns of nucleotide substitution have stimulated the development of more elaborate methods (e.g., Holmquist and Pearl 1980; Kimura 1981; Gojobori et al. 1982a; Kaplan and Risko 1982; Lanave et al. 1984; Tajima and Né 1984).

A different approach has also been taken in order to estimate the number of synonymous substitutions (which cause no amino acid replacement) and that of
nonsynonymous substitutions separately (Kimura 1980; Miyata and Yasunaga 1980; Perler et al. 1980). This approach is desirable when the amino acid coding regions of genes are compared, for it is known that the rate of synonymous substitution is generally much greater than that of nonsynonymous substitution (e.g., Kafatos et al. 1977; Kimura 1977; Jukes and King 1979; Miyata et al. 1980). Currently, the most frequently used methods are those proposed by Miyata and Yasunaga (1980) and Perler et al. (1980). There are, however, several problems with these two methods. They are discussed fully later on, but for the present we will note that they involve the distribution of substitutions among sites, correction for multiple hits, and the weighting of alternative paths between codon pairs with more than one nucleotide difference. Because of these three problems, the two methods tend to give biased estimates when the degree of sequence divergence becomes relatively large (Brown et al. 1982; Miyata et al. 1982; Gojobori 1983). In an attempt to get better estimates, we have developed a new method. It is based on the pattern of codon degeneracy in the codon table and adopts Kimura’s (1980) method of distinguishing between transitional and transversional substitutions. It therefore allows for unequal rates of substitution among nucleotides caused by either differences between transitional and transversional rates or differences between synonymous and nonsynonymous rates. Correction for multiple substitutions can also be done in a more rigorous manner. Moreover, the relative likelihood of synonymous and nonsynonymous substitutions is determined from codon changes in mammalian genes and should therefore be more reliable than that determined from protein sequence data. As will be seen later, our method also provides more information about the substitution rates at nucleotide sites with different degrees of codon degeneracy.

After describing the method, we shall apply it to a large number of mammalian genes and compare the synonymous rate with the nonsynonymous rate in the same gene and synonymous (or nonsynonymous) rates among different genes. In addition, we shall discuss phylogenetic reconstruction and the implication of the relative likelihood of codon changes for the mechanisms of DNA evolution.

Method

In presenting the method, we shall use the nuclear genetic code as an example. Although the method is developed primarily for the study of nuclear genes, the regularity in codon degeneracy in the mammalian mitochondrial (mt) genetic code (Anderson et al. 1981) makes it even more suitable for studying mammalian mt genes.

We classify nucleotide sites into nondegenerate, twofold degenerate, and fourfold degenerate sites. A site is fourfold degenerate if all possible changes at the site are synonymous. The third positions of 32 of the 61 sense codons—e.g., GTN (Val), N being any base—are of this type. A site is twofold degenerate if one of the three possible changes is synonymous. The third positions of 24 of the 61 sense codons—e.g., CAY (His) (Y = T or C)—and the first positions of four leucine codons YAR (R = A or G) and four arginine codons CGR and AGR are of this type. (In mammalian mt genes, AGR are stop codons [Anderson et al. 1981] and hence the first positions of CGR [Arg] are nondegenerate.) We also include the third position of the three isoleucine codons in this class, though they are actually threefold degenerate sites. (In mammalian mt genes, there are only two codons for isoleucine and their third position is indeed a twofold degenerate site.) A site is
nondegenerate if all possible changes at this site are nonsynonymous or nonsense. The second positions of all sense codons and the first positions of most codons belong to this class, and so do the third positions of ATG (Met) and TGG (Trp). (In mammalian mt genes, there are two codons for Met and Trp, and their third positions are twofold degenerate sites.) We count the numbers of sites in these three classes in each of the two sequences compared and then compute the average numbers, denoting them by \( L_0 \) (nondegenerate), \( L_2 \) (twofold), and \( L_4 \) (fourfold), respectively.

The purpose of such classifications is to enable one to estimate synonymous and nonsynonymous rates of nucleotide substitution separately. Substitutions at nondegenerate sites and fourfold degenerate sites are nonsynonymous and synonymous, respectively. At the twofold degenerate sites, transversions (purine-pyrimidine changes) lead to nonsynonymous changes, whereas transitions (C \( \leftrightarrow \) T or A \( \leftrightarrow \) G) lead to synonymous changes. There is no exception to this rule in the mammalian mt genetic code. In the nuclear genetic code, there are two exceptions: the first positions of CGR and AGR (all for Arg) and the last positions of ATH (all for Ile, H = not-G). Here, we make an ad hoc adjustment to overcome this minor irregularity: we classify all synonymous changes in these two cases as "transitions" and all nonsynonymous changes as "transversions." For example, CGG (Arg) \( \leftrightarrow \) AGG (Arg) is considered "transition," whereas CGG (Arg) \( \leftrightarrow \) TGG (Trp) is considered "transversion," though the contrary is actually true. With these modifications, the aforementioned rule about the twofold degenerate sites also applies to the nuclear genetic code.

We then compare two sequences codon by codon and classify each difference as either transition or transversion. We denote \( P_i \) and \( Q_i \) as the number of observed transitional and transversional differences, respectively, at \( i \)-fold degenerate sites divided by \( L_i \) (the total number of \( i \)-fold degenerate sites; \( i = 0, 2, \text{ or } 4 \)). For example, suppose that the two sequences compared are TTT CTA and TCT CTG, respectively. In the first sequence, the first two positions of TTT (Phe) are nondegenerate sites and the third position is a twofold degenerate site; the first position of CTA (Leu) is a twofold degenerate site, the second position is a nondegenerate site, and the third position is a fourfold degenerate site. Therefore, in the first sequence, \( L_0 = 2 + 1 = 3 \), \( L_2 = 1 + 1 = 2 \), and \( L_4 = 0 + 1 = 1 \). Similarly, we have \( L_0 = 3 \), \( L_2 = 1 \), and \( L_4 = 2 \) for the second sequence. From these, we obtain the averages \( L_0 = (3 + 3)/2 = 3 \), \( L_2 = 1.5 \), and \( L_4 = 1.5 \). Furthermore, we note that the codon pair TTT and TCT have a transitional difference at a nondegenerate site and that the pair CTA and CTG have a transitional difference at a fourfold degenerate site. Therefore, \( P_0 = 1/3 \), \( P_4 = 1/1.5 \) and the other four \( P_i \) and \( Q_i \) values are 0.

Kimura (1980) developed a two-parameter method to estimate the number of transitional and transversional substitutions per site based on the observed proportions of transitional and transversional differences. To apply his method, we assume that the transitional and transversional rates of substitution are \( \alpha_i \) and \( 2\beta_i \) substitutions per \( i \)th type site per unit time, respectively. Using his formulas 8 and 9, we can readily show that the mean and approximate error variance of the number of transitional (\( A_i \)) and transversional (\( B_i \)) substitutions per \( i \)th type site are given by

\[
A_i = (1/2) \ln(a_i) - (1/4) \ln(b_i),
\]

\[
V(A_i) = \frac{a_i^2 P_i + c_i^2 Q_i - (a_i P_i + c_i Q_i)^2}{L_i},
\]
\[ B_i = \frac{1}{2} \ln (b_i), \quad (3) \]
\[ V(B_i) = b_i^2 Q_i (1 - Q_i)/L_i, \quad (4) \]

where \( a_i = 1/(1 - 2P_i - Q_i) \), \( b_i = 1/(1 - 2Q_i) \), and \( c_i = (a_i - b_i)/2 \). The total number \((K_i)\) of substitutions per \(i\)th type site is given by
\[ K_i = A_i + B_i, \quad (5) \]

with a variance approximately equal to
\[ V(K_i) = [a_i^2 P_i + d_i^2 Q_i - (a_i P_i + d_i Q_i)^2]/L_i, \quad (6) \]

where \( d_i = b_i + c_i \). We note that \( A_2 \) and \( B_2 \) denote the number of synonymous and nonsynonymous substitutions per twofold degenerate site, \( K_4 = A_4 + B_4 \) the number of synonymous substitutions per fourfold degenerate site, and \( K_0 = A_0 + B_0 \) the number of nonsynonymous substitutions per nondegenerate site.

Let \( K_S \) be the number of (synonymous) substitutions per synonymous site and \( K_N \) the number of (nonsynonymous) substitutions per nonsynonymous site. Following convention, we count each fourfold degenerate site as a synonymous site, each nondegenerate site as a nonsynonymous site, and each twofold degenerate site as one-third synonymous and two-thirds as nonsynonymous. (This is because, if a twofold degenerate site mutates randomly, only one-third of the time would it be scored as synonymous and two-thirds of the time as nonsynonymous.) We then obtain
\[ K_S = (L_2 A_2 + L_4 K_4)/(L_2/3 + L_4) = 3(L_2 A_2 + L_4 K_4)/(L_2 + 3L_4), \quad (7) \]
\[ V(K_S) = 9[L_2^2 V(A_2) + L_4^2 V(K_4)]/(L_2 + 3L_4)^2, \quad (8) \]
\[ K_A = (L_2 B_2 + L_0 K_0)/(2L_2/3 + L_0) = 3(L_2 B_2 + L_0 K_0)/(2L_2 + 3L_0), \quad (9) \]
\[ V(K_A) = 9[L_2^2 V(B_2) + L_0^2 V(K_0)]/(2L_2 + 3L_0)^2. \quad (10) \]

There are also cases where nucleotide differences cannot be unambiguously classified. This happens when two codons differ by more than one nucleotide; for example, CCC (Pro) versus CAA (Gln). The change at the second position can be unambiguously included in \( Q_0 \) but that at the third position may be included in either \( Q_4 \) or \( Q_2 \), depending on whether the pathway of changes was CCC \( \leftrightarrow \) CCA \( \leftrightarrow \) CAA or CCC \( \leftrightarrow \) CAC \( \leftrightarrow \) CAA. (Note that there are one synonymous and one nonsynonymous change in the former pathway but that there are two nonsynonymous changes in the latter.) Since \( Q_4 \) is used to estimate the synonymous rate and \( Q_2 \) is used for the nonsynonymous rate, misclassification may result in erroneous estimates. The situation is more complex, though much rarer, when two compared codons differ at all three positions. For example, among the 6 (=3!) pathways between TCG (Ser) and CGT (Arg), three pathways require three nonsynonymous changes, one pathway requires two, and the remaining two pathways require only one nonsynonymous change. To determine the probabilities of different pathways in ambiguous cases, adequate information on the relative likelihood of codon changes is necessary. This is the subject of the following section.
A computer program for our method is available by sending us a floppy disk. Although the computation algorithm is somewhat complicated, the program is very easy to use.

Relative Likelihood of Codon and Amino Acid Changes

In the method of Perler et al. (1980), all codon changes are implicitly assumed to be equally probable. Miyata and Yasunaga (1980), on the other hand, determine the relative likelihood of codon changes by extrapolating from previous analyses of amino acid exchangeability (McLachlan 1971; Grantham 1974; Miyata et al. 1979). In this section, we evaluate such likelihoods from DNA sequence data by taking the following steps: (1) The codons at all ancestral nodes of a phylogenetic tree are inferred from the present-day DNA sequences, and the frequencies of codon changes between every node and its immediate descendant are recorded as observed frequencies. (2) The expected frequencies of codon changes are formulated. (3) The observed and expected codon changes are classified according to the differences in the physicochemical properties between the encoded amino acids. The relative likelihood of a particular codon or amino acid change is simply the ratio of the observed frequency to the expected one.

Inferring the Ancestral Codons

Fitch (1971) developed an algorithm that, given a phylogenetic tree, will find all the minimum distance solutions at any nucleotide position. (A minimum distance solution is an assignment to every ancestral node of a nucleotide that gives the smallest number of nucleotide substitutions for the given tree.) If the DNA sequences available constitute an adequately branched tree, the nucleotide at each ancestral node can be uniquely determined; of course, uniqueness does not guarantee correctness (Holmquist 1979). The ancestral codon can certainly be uniquely determined if all three positions are. When the codon at any of the ancestral nodes has more than one solution, we take the average by assigning a probability to each solution as Fitch (1971) did and further assume that all positions are independent of one another. Therefore, if T has a probability of 1/4 at the first position, C is uniquely determined at the second position, and A has a probability of 1/4 at the third position, and the probability of having TCA is 1/16. Rules for connecting an ancestral codon to its immediate descendants can be similarly extended from the description of Fitch (1971). If the codons of adjacent nodes differ at more than one position, the "average" solution, i.e., each position assumed to have an equal probability of substitution prior to others, is used. This procedure favors nonsynonymous over synonymous substitutions. It will, however, introduce no serious bias, for the proportion of cases where the ancestral codon cannot be uniquely determined is small in the data we used.

Once the ancestral codons have been assigned, we can then record the frequencies of observed changes between any pair of codons differing by one nucleotide.

Expected Frequencies of Codon and Amino Acid Changes

If codon changes are determined solely by mutation, then the relative frequencies of changes from, say, TCA to CCA, ACA, GCA, TTA, TC7, TCC, and TCG should be $f_{TC}$, $f_{TA}$, $f_{TG}$, $f_{CT}$, $f_{AT}$, $f_{AC}$, and $f_{AG}$, respectively, where $f_{XY}$ is the relative
frequency of substituting nucleotide Y for X as defined by Gojobori et al. (1982b). Substitution of A or G for C at the second position of TCA results in one of the termination codons and is not counted.

Let us now number the codons from 1 to 61, denoted by n, by varying the third position first, followed by the second and then the first position. The three termination codons are skipped. We also label the amino acids, in alphabetical order, from 1 to 20. We may now define \( P_n(j) \) to be the probability that the \( n \)th codon, after one nucleotide substitution, codes for the \( j \)th amino acid (denoted aa\(_j\)). In the above example, TCA is the seventh codon, CCA codes for Pro (aa\(_{15}\)), ACA codes for Thr (aa\(_{17}\)), and so on. Therefore, \( P_7(15) = f_{TC} \), \( P_7(17) = f_{TA} \), and so on. \( P_n(j) = 0 \) if the \( n \)th codon cannot change to a codon for aa\(_j\) in one step. The expected relative frequency of having aa\(_i\), replaced by aa\(_j\), in the genes studied can then be expressed as

\[
e_{ij} = \sum_{n \in N_i} P_n(j) T_n,
\]

where \( T_n \) is the observed number of the \( n \)th codon in the inferred ancestral sequences and \( N_i \) is the set of codons coding for the \( i \)th amino acid. The quantity \( P_n(j) T_n \) is summed over all codons that code for aa\(_i\). Note that \( e_{ij} \) (the expected frequency of synonymous changes between codons for aa\(_i\)) is well defined and that \( e_{ij} \neq e_{ji} \). Similarly, we can calculate the observed frequency (\( o_{ij} \)) of changes from aa\(_i\) to aa\(_j\) from the changes between codons inferred from DNA sequences.

Classification of Codon Changes and Their Relative Likelihoods

Since there would be 400 elements in the \( [e_{ij}]_{20 \times 20} \) matrix, each observed element \( o_{ij} \) may be subject to large sampling errors. We therefore classify all the amino acid changes into four classes on the basis of Grantham's (1974) distance between amino acids. He assigns a distance, \( d_{ij} \), between aa\(_i\) and aa\(_j\) according to the physicochemical properties of the two amino acids. The value of \( d_{ij} \) ranges from 5 (Ile-Leu) to 215 (Cys-Trp). Li et al. (1984) define amino acid changes as conservative, moderately conservative, moderately radical, and radical if \( d_{ij} \) is less than 50, between 50 and 100, between 100 and 150, and greater than 150, respectively. Using this definition, we sum up all \( e_{ij} \)'s and \( o_{ij} \)'s with \( d_{ij} \)'s in the same range. In addition, we have the synonymous class in which \( e_{ij} \)'s and \( o_{ij} \)'s are summed from \( i = 1, 2, \ldots, 20 \).

Table 1 gives the results of the relative likelihood of codon changes for genes with moderate rates of substitution and for those with very high rates (see below). Conservative genes like histone and actin genes have too few amino acid replacements for such a survey. The observed changes are not whole numbers because they represent the average over all possible ancestral codons. The expected changes are normalized so that the total number of expected changes equals that of observed ones. Since only mammalian genes are used, multiple substitutions beyond the minimum distance solution should not be a serious problem for nonsynonymous substitutions. This assumption, however, tends to underestimate to some extent the number of synonymous substitutions because the chance of multiple synonymous substitutions at the same site may not be negligible between sequences from different mammalian orders (e.g., see Holmquist et al. 1982). In the case of genes evolving
Table 1
Relative Likelihoods of Codon Changes*

<table>
<thead>
<tr>
<th>Types of Changes</th>
<th>Genes with Moderate Substitution Ratesb</th>
<th>Genes with High Substitution Ratesc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymous</td>
<td>960.5/429.8 = 2.235</td>
<td>477.6/292.1 = 1.635</td>
</tr>
<tr>
<td>Conservative</td>
<td>378.5/443.7 = 0.853</td>
<td>331.3/318.2 = 1.041</td>
</tr>
<tr>
<td>Moderately conservative</td>
<td>401.1/522.8 = 0.767</td>
<td>337.8/375.8 = 0.899</td>
</tr>
<tr>
<td>Moderately radical</td>
<td>89.2/313.1 = 0.285</td>
<td>108.5/222.7 = 0.487</td>
</tr>
<tr>
<td>Radical</td>
<td>11.7/131.5 = 0.089</td>
<td>36.8/83.2 = 0.442</td>
</tr>
<tr>
<td>Total</td>
<td>1841</td>
<td>1292</td>
</tr>
</tbody>
</table>

*The relative likelihood is the ratio of the observed frequency to the expected one. The expected frequencies are calculated according to the description in the text, with the assumption that substitution rates among all nucleotide pairs are equal.


at moderate rates, we choose only those for which DNA sequences from three or more mammalian species have been available. This helps to reduce ambiguity in inferring ancestral codons. For very rapidly evolving genes, we have to use also genes for which only two sequences are available.

From table 1, we find that the greater the physicochemical distance between two amino acids is, the less likely one will replace the other. For example, for every moderately radical replacement in a gene with a moderate nonsynonymous rate, 7.84 (=2.235/0.285) synonymous changes would be observed. (If there were no selective constraints on amino acid replacement, the ratio would be 1.) Not surprisingly, in rapidly evolving genes such as relaxin and immunoglobulin genes, amino acid exchanges are accepted to a much greater extent; even radical replacements occur at a substantial rate.

An interesting point from table 1 is that one should not extrapolate the relative likelihood (obs/exp) of amino acid replacement to estimate that of synonymous changes as suggested by Miyata and Yasunaga (1980). If we draw a regression line for the relative likelihood of amino acid exchanges on the amino acid distance (dij, using the midpoint of each class), we obtain good correlation coefficients for both types of genes (r = −0.97 and −0.96, respectively). However, the intercepts at dij = 0 are 1.05 and 1.16 for the two types of genes, respectively. These extrapolated values are considerably lower than the likelihoods (2.24 and 1.64, respectively) shown in table 1.

To apply the results of table 1, we should consider genes with slow, moderate, and high nonsynonymous rates separately. The distinctions are certainly not well
defined, but, generally speaking, this means genes with less than 0.6, between 0.6 and 1.2, and greater than 1.2 nonsynonymous substitutions per nonsynonymous site per 10^9 years, respectively. Table 1 does not contain the relative likelihood of codon changes in slowly evolving genes. To estimate the substitution rates in conservative genes like histone, actin, and many hormone genes, we may use the relative likelihoods derived from “moderately conservative” genes. This is because conservative genes rarely require probability assignments to different substitution pathways, so that the results depend little on how we weight the pathways. The exceptions are codons for Leu and Arg. The two pathways between TTA and CTT (both for Leu) may pass through either CTA (Leu) or TTT (Phe). The former takes two synonymous substitutions, whereas the latter requires two exchanges between Leu and Phe. In such cases, we assume that two synonymous substitutions have taken place.

Finally, the relative likelihood of each pathway is simply the product of the relative likelihoods of the codon changes along the pathway (see Miyata and Yasunaga [1980] for details). For example, consider the pathway between CCC and CAA via CCA. Since CCC (Pro) ↔ CCA (Pro) is synonymous with a relative likelihood \( r = 2.235 \) and since CCA (Pro) ↔ CAA (Gln) is moderately conservative, with \( r = 0.767 \), this pathway is given a relative likelihood of \( 2.235 \times 0.767 = 1.71 \). Similarly, the other pathway, CCC (Pro) ↔ CAC (His) ↔ CAA (Gln), is given the relative likelihood of \( 0.767 \times 0.853 = 0.654 \). The probability of the former pathway is therefore \( 1.71/(1.71 + 0.654) = 0.723 \) and that of the latter is 0.277. In other words, between CCC and CAA, we assume that 0.723 synonymous changes and 0.723 + 2 \times 0.277 = 1.276 nonsynonymous changes have occurred. Codons differing in all three positions can be similarly treated.

**Substitution Rates in Mammalian Genes**

In this section, we apply our method to study the rates of synonymous and nonsynonymous substitution in various kinds of genes. For several reasons, we shall use mainly genes from mammals: man, rodents (mouse, rat, and Chinese hamster), and artiodactyls (cow, goat, and pig). First, the fossil record for mammals is better than those for other organisms; we assume that the above mammalian orders diverged 80 Myr ago. Second, many genes from mammals have been sequenced. Third, to get a reliable estimate of the rate of nucleotide substitution, the degree of sequence divergence should not be too small or too large. For this reason, genes from different mammalian orders are generally very suitable for estimating both the synonymous and nonsynonymous rates. For extremely conservative genes, such as those coding for histones and glucagon, estimates of nonsynonymous rates from comparisons of mammalian species may not be reliable, for the number of nonsynonymous substitutions is too small; and we have therefore also obtained estimates from comparisons between mammals and chicken or fish.

The results of our analysis are presented in table 2. In those cases where more than two sequences are used, the mean (\( d \)) and variance (\( V(d) \)) are obtained as follows. Suppose the \( n \) sequences used are from \( n \) mammalian orders. We assume that all mammalian orders diverged at the same time, i.e., 80 Myr ago. For each type of nucleotide site, we first compute the mean (\( \bar{d}_{ij} \)) and variance (\( V(d_{ij}) \)) of the number of substitutions between sequences \( i \) and \( j \). We then compute \( d \) and \( V(d) \) according to the following formulas:
<table>
<thead>
<tr>
<th>GENES</th>
<th>NONDEGENERATE SITES</th>
<th>TWOFOLD DEGENERATE SITES</th>
<th>FOURFOLD DEGENERATE SITES</th>
<th>NON-SYNONYMOUS RATE</th>
<th>SYNONYMOUS RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L</td>
<td>Transition</td>
<td>Transversion</td>
<td>Total (±SE)</td>
<td>Transition</td>
</tr>
<tr>
<td>Histones:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H4</td>
<td>101</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00 (0.00)</td>
<td>1.99</td>
</tr>
<tr>
<td>H3</td>
<td>101</td>
<td>0.00&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.00</td>
<td>0.00 (0.00)</td>
<td>0.43</td>
</tr>
<tr>
<td>H3</td>
<td>135</td>
<td>0.00</td>
<td>0.03</td>
<td>0.00 (0.00)</td>
<td>2.33</td>
</tr>
<tr>
<td>H2B</td>
<td>120</td>
<td>0.03&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.04</td>
<td>0.07 (0.04)</td>
<td>1.44</td>
</tr>
<tr>
<td>H2A</td>
<td>126</td>
<td>0.02&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.06</td>
<td>0.08 (0.03)</td>
<td>0.75</td>
</tr>
<tr>
<td>Contractile system proteins:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actin, α</td>
<td>376</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01 (0.01)</td>
<td>1.81</td>
</tr>
<tr>
<td>Hormones and active peptides:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Somatostatin-28</td>
<td>28</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00 (0.00)</td>
<td>0.52</td>
</tr>
<tr>
<td>Gastrin</td>
<td>82</td>
<td>0.00</td>
<td>0.09</td>
<td>0.09 (0.12)</td>
<td>0.78</td>
</tr>
<tr>
<td>Glucagon</td>
<td>29</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00 (0.00)</td>
<td>3.42</td>
</tr>
<tr>
<td>Insulin</td>
<td>5</td>
<td>0.02</td>
<td>0.09</td>
<td>0.11 (0.07)</td>
<td>2.34</td>
</tr>
<tr>
<td>ACTH</td>
<td>39</td>
<td>0.05</td>
<td>0.16</td>
<td>0.22 (0.11)</td>
<td>1.06</td>
</tr>
<tr>
<td>Parathyroid hormone</td>
<td>90</td>
<td>0.34</td>
<td>0.18</td>
<td>0.52 (0.15)</td>
<td>0.89</td>
</tr>
<tr>
<td>Corticotropin releasing factor</td>
<td>41</td>
<td>0.32</td>
<td>0.34</td>
<td>0.66 (0.13)</td>
<td>0.75</td>
</tr>
<tr>
<td>Glycoprotein hormone, α</td>
<td>92</td>
<td>0.36</td>
<td>0.42</td>
<td>0.78 (0.14)</td>
<td>2.99</td>
</tr>
<tr>
<td>Growth hormone</td>
<td>189</td>
<td>0.67</td>
<td>0.35</td>
<td>1.02 (0.12)</td>
<td>2.28</td>
</tr>
<tr>
<td>Insulin C-peptide</td>
<td>35</td>
<td>0.29</td>
<td>0.75</td>
<td>1.04 (0.29)</td>
<td>1.51</td>
</tr>
<tr>
<td>Prolactin</td>
<td>197</td>
<td>0.77</td>
<td>0.57</td>
<td>1.34 (0.14)</td>
<td>1.84</td>
</tr>
<tr>
<td>Relaxin C-peptide</td>
<td>99</td>
<td>1.19</td>
<td>0.85</td>
<td>2.03 (0.27)</td>
<td>1.79</td>
</tr>
<tr>
<td>Relaxin</td>
<td>54</td>
<td>1.51</td>
<td>1.30</td>
<td>2.81 (0.44)</td>
<td>2.54</td>
</tr>
<tr>
<td>Signal Peptides:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Somatostatin</td>
<td>23</td>
<td>0.00</td>
<td>0.31</td>
<td>0.31 (0.23)</td>
<td>0.70</td>
</tr>
<tr>
<td>Insulin</td>
<td>23</td>
<td>0.65</td>
<td>0.56</td>
<td>1.21 (0.39)</td>
<td>0.44</td>
</tr>
<tr>
<td>Relaxin</td>
<td>23</td>
<td>1.87</td>
<td>0.63</td>
<td>2.50 (0.71)</td>
<td>2.33</td>
</tr>
<tr>
<td>Hemoglobins:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>α chain</td>
<td>141</td>
<td>0.19</td>
<td>0.40</td>
<td>0.59 (0.10)</td>
<td>1.97</td>
</tr>
<tr>
<td>β chain</td>
<td>144</td>
<td>0.35</td>
<td>0.53</td>
<td>0.89 (0.13)</td>
<td>1.48</td>
</tr>
<tr>
<td>Protein</td>
<td>Value</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immunoglobulin (Ig) and related proteins:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ig Vκ</td>
<td>0.365</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ig κ</td>
<td>0.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ig γ</td>
<td>0.365</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ig κ</td>
<td>0.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β2-microglobulin</td>
<td>0.41</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ig γ</td>
<td>0.365</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Interferons:**

<table>
<thead>
<tr>
<th>Protein</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>α1</td>
<td>0.07</td>
</tr>
<tr>
<td>α2</td>
<td>0.05</td>
</tr>
<tr>
<td>β1</td>
<td>0.04</td>
</tr>
<tr>
<td>γ</td>
<td>0.03</td>
</tr>
</tbody>
</table>

**Other proteins:**

<table>
<thead>
<tr>
<th>Protein</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPRT</td>
<td>0.33</td>
</tr>
<tr>
<td>Neurotensin II</td>
<td>0.33</td>
</tr>
<tr>
<td>Metallothionein II</td>
<td>0.33</td>
</tr>
<tr>
<td>Fibrinogen γ</td>
<td>0.30</td>
</tr>
<tr>
<td>Metallothionein</td>
<td>0.30</td>
</tr>
<tr>
<td>Albumin</td>
<td>0.28</td>
</tr>
<tr>
<td>a-Lactalbumin</td>
<td>0.25</td>
</tr>
<tr>
<td>α-Fetoprotein</td>
<td>0.25</td>
</tr>
<tr>
<td>Average^d</td>
<td>0.48</td>
</tr>
</tbody>
</table>

* Unless stated otherwise, all rates are obtained from comparisons between different mammalian orders, which are assumed to have diverged 80 Myr ago. All rates are in substitutions per site per 10⁸ year.

* References: histone H1 (Zhong et al. 1983; Seiler-Tuyns and Birstein 1981; Sagarman et al. 1983); histone H3 (Zhong et al. 1983; Sittman et al. 1983; Engel et al. 1982); histone H2A (Zhong et al. 1983; Harvey et al. 1982); histone H2B (Zhong et al. 1983; Sittman et al. 1983); actin-a (Hanauer et al. 1983; Zakut et al. 1982); somatostatin-28 (Shen et al. 1982; Goodman et al. 1983); gastrin (Kato et al. 1983; Yoo et al. 1982); glucagon (Bell et al. 1983b; Lopez et al. 1983); insulin (Bell et al. 1980; Wetskam et al. 1982; Kwok et al. 1983; Lomdock et al. 1979); ACTH (Takahashi et al. 1983; Nakashita et al. 1981; Notake et al. 1983); parathyroid hormone (Hendy et al. 1981); corticotropin releasing factor (Shibahara et al. 1983); glycoprotein hormone, α subunit (Fiddes et al. 1981; Nilson et al. 1983; Chin et al. 1981; Godine et al. 1982); growth hormone (Miller and Eberhardt 1983; Seeburg et al. 1983); prolactin (Miller and Eberhardt 1983); relaxin (Hudon et al. 1983; Hudson et al. 1981); α-antitrypsin (Hudon et al. 1982); hemoglobin (Hudon et al. 1980); haptoglobin (Hudon et al. 1980); γ-globulin (Lawn et al. 1980); Schon et al. 1981; Hardison et al. 1979; Koekel et al. 1979; immunoglobulin (Ig) Vκ (Rechavi et al. 1982); Ig κ (Hieter et al. 1981; Bothwell et al. 1982); β2-microglobulin (Saug et al. 1981; Parney and Seidman 1982); Ig γ1 (Ellison et al. 1982); Ig κ (Hieter et al. 1980); Sheppard and Gutman 1981; Emonor et al. 1983); interferon-α2 (Shaw et al. 1983); interferon-α (Lawn et al. 1981); Shaw et al. 1983); interferon-β1 (Higashi et al. 1983); interferon-γ (Gray and Goeddel 1983); HPRT (Jolly et al. 1983); Konecki et al. 1982); neurophysin II (Laid et al. 1982; Schmaltz et al. 1983); metallothionein II (Karin and Richards 1982; Schmidt and Hauer 1983; Griffith et al. 1983); metallothionein I (Andersen et al. 1983; Schmidt and Hauer 1983; Griffith et al. 1983; Fibrinogen γ (Chung et al. 1983); Crabtree and Kast 1982); Albumin (Dugaicyzk et al. 1982); Sargent et al. 1981); α-lactalbumin (Hall et al. 1982; Dandecker et al. 1981); and a-fetoprotein (Morinaga et al. 1983; Law and Dugaicyzk 1981; Jagudzinski et al. 1981).

^ Number of codons compared.

^ All rates in these cases are obtained from comparisons between mammals and birds or fish. The divergence time between mammals and birds is assumed to be 270 Myr and that between mammals and fish to be 400 Myr.

^ In this case, the mean and variance of the rate of nonsynonymous substitution are obtained from comparisons between mammals and birds or fish.

^ The average for all proteins is obtained only from comparisons between mammalian orders.
\[ \bar{d} = \sum_{i < j}^{n} \bar{d}_{ij}/[n(n-1)/2], \]

\[ V(d) = 4 \sum_{i < j}^{n} V(d_{ij})/[n^2(n-1)]. \]

In the case where more than one sequence from the same mammalian order are used, formulas for computing \( \bar{d} \) and \( V(d) \) can also be obtained (see Wu and Li 1985). For simplicity, however, we have computed \( \bar{d} \) and \( V(d) \) by taking a simple average. For example, if sequence 1 is from man and sequences 2 and 3 are from mouse and rat, then the relations \( \bar{d} = (d_{12} + d_{13})/2 \) and \( V(d) = [V(d_{12}) + V(d_{13})]/2 \) are used. In each gene group of table 2, the results are arranged in the increasing order of the rate of nonsynonymous substitution. In the following we discuss a number of interesting features.

First, the rate of nonsynonymous substitution is extremely variable among different kinds of genes; it ranges from \( 0.004 \times 10^{-9} \) (histone H4) to 2.80 \( \times 10^{-9} \) (interferon-\( \gamma \)) substitutions per nonsynonymous site per year, a more than hundredfold difference (see the next to last column of table 2). As is well known, histone and actin genes are extremely conservative. Some hormone genes, e.g., somatostatin-28, glucagon, gastrin, and insulin genes, are also extremely conservative, but some others, e.g., relaxin and prolactin, evolve rapidly. The insulin C peptide has often been used as an example of rapid evolution, but it actually evolves considerably more slowly than relaxin and the relaxin C peptide. The nonsynonymous rate observed in the signal peptide region of the insulin gene (1.2 \( \times 10^{-9} \)) is about the average for signal peptide regions (results not shown); the extremely high and low rates in the signal peptide regions of the relaxin and somatostatin genes, respectively, are probably extreme random deviates, for signal peptide regions are very short. Hemoglobin genes evolve at an intermediate rate, while immunoglobulin and interferon genes evolve rapidly. The average rate of nonsynonymous substitution for all the gene comparisons from mammals is 0.88 \( \times 10^{-9} \) substitutions per nonsynonymous site per year (see the last row). Interestingly, the median of the nonsynonymous rates in table 2, 0.87 \( \times 10^{-9} \) for \( \beta \)-globin, is almost identical with the mean.

Second, the rate of synonymous substitution varies greatly from gene to gene (see the last column of table 2), though not so much as the rate of nonsynonymous substitution. The variation is especially great among genes with fewer than 50 fourfold degenerate sites; for example, in \( \beta 2 \)-microglobulin the rate is as high as 11.8 \( \times 10^{-9} \), while in parathyroid hormone it is as low as 1.7 \( \times 10^{-9} \). It appears that our results do not support Miyata et al.'s (1980) suggestion, based on a considerably smaller number of genes, that the rate of synonymous substitution is similar among genes. From table 2, the average rate of synonymous substitution for all the comparisons of genes from mammals is 4.7 \( \times 10^{-9} \) substitutions per synonymous site per year. This is five times higher than the average rate of nonsynonymous substitution. Our estimate is not much different from Hayashida and Miyata's (1983) estimate of 5.5 \( \times 10^{-9} \). (Their estimate reduces to 5.1 \( \times 10^{-9} \) if 80 Myr instead of 75 Myr is used as the divergence time between mammalian orders. Note also that we used 36 genes, while they used only 11 genes.) Interestingly, in histones H4 and H3 the synonymous rates are very high, though the nonsynon-
yinous rates are extremely low. There is, however, a tendency for a gene with a high nonsynonymous rate to have a high synonymous rate; in the majority of genes with a nonsynonymous rate higher than $1.0 \times 10^{-9}$, the synonymous rate is higher than the average.

To examine the issue of uniform synonymous rate in greater detail, we have used only those genes in table 2 for which there are more than 50 fourfold degenerate sites (to reduce stochastic noise) and compared only sequences from different mammalian orders. Another source of variation may be introduced if DNA sequences from rodents are used in the studies of some genes but not in others. There is evidence that synonymous rates as well as substitution rates in weakly constrained DNA regions are significantly higher in both mouse and rat than in other mammals (Wu and Li 1985). One plausible explanation for this is the very short generation time of these rodents. For each gene, we therefore separate the comparisons of DNA sequences that involve a rodent sequence from those that do not. We then compute the averages for these two types of comparisons separately.

(A gene may thus have two different synonymous rates, one involving rodents and the other involving only other mammalian orders.)

There are 21 genes from table 2 that meet the above criteria and have at least one rodent sequence. The synonymous rates of these genes involving rodent sequences are distributed as follows: one less than $2.5 \times 10^{-9}$, eight between $2.5 \times 10^{-9}$ and $5.0 \times 10^{-9}$, and 12 between $5.0 \times 10^{-9}$ and $7.5 \times 10^{-9}$; the mean of these rates is $4.97 \times 10^{-9}$. There are only 10 genes that have sequences from two nonrodent mammalian orders; the distribution of their synonymous rates are as follows: three less than $2.5 \times 10^{-9}$, six between $2.5 \times 10^{-9}$ and $5.0 \times 10^{-9}$, and one between $5.0 \times 10^{-9}$ and $7.5 \times 10^{-9}$; the mean is $3.41 \times 10^{-9}$. Therefore, the synonymous rates in comparisons that involve rodents tend to be considerably higher than those in comparisons that do not. In both sets of comparisons, however, synonymous rates among different genes do not appear to be as uniform as suggested by Miyata et al. (1980) and Hayashida and Miyata (1983); the highest rates are about three times the lowest ones in both sets. Caution therefore should be used when synonymous rates are used as a molecular clock to date the divergence time between genes or species.

Third, at fourfold degenerate sites the rate of transitional substitution tends to be higher than the rate of transversional substitution, though at each site two types of transversional change and only one type of transitional change can occur. The averages for these two rates are $2.5 \times 10^{-9}$ and $1.7 \times 10^{-9}$, respectively. This observation can be explained by the fact that transitional mutation occurs more frequently than transversional mutation (Li et al. 1984). At twofold degenerate sites, the rate of transitional substitution is on the average about the same as that at fourfold degenerate sites, but the rate of transversional substitution is usually considerably lower than the corresponding rate at fourfold degenerate sites; the averages for the two rates at twofold degenerate sites are $1.8 \times 10^{-9}$ and $0.44 \times 10^{-9}$, respectively. This lower rate of transversional substitution occurs because all transversional changes at twofold degenerate sites are nonsynonymous. At nondegenerate sites, the rates of transitional and transversional substitution are on the average about the same ($0.48 \times 10^{-9}$ and $0.45 \times 10^{-9}$, respectively) and are usually considerably lower than the corresponding values at fourfold degenerate sites. These low rates occur because all changes at nondegenerate sites are nonsyn-
nonymous. Thus, the (total) rate of nucleotide substitution is lowest at nondegenerate sites, intermediate at twofold degenerate sites, and highest at fourfold degenerate sites; the averages are $0.94 \times 10^{-9}$, $2.3 \times 10^{-9}$, and $4.2 \times 10^{-9}$, respectively.

Fourth, the rate of nucleotide substitution at fourfold degenerate sites tends to be lower than the rate of synonymous substitution (last column in table 2), though all changes at fourfold degenerate sites are synonymous. Indeed, the average value is only $4.2 \times 10^{-9}$ for the former rate but $4.7 \times 10^{-9}$ for the latter rate. This difference occurs because in computing the rate of synonymous substitution, the transitional changes at twofold degenerate sites are also included. As mentioned above, only one-third of a twofold degenerate site is counted as synonymous but transitional (synonymous) substitution at a twofold degenerate site occurs at a rate higher than one-third of the total rate at a fourfold degenerate site. For this reason, the synonymous rate obtained by the conventional definition tends to be higher than the rate of nucleotide substitution at fourfold degenerate sites. In inferring the stringency of functional constraint on synonymous changes, it has been customary to compare the rate of synonymous substitution with the substitution rate in pseudogenes, which is taken as the neutrality standard. It now appears that for this purpose one should compare instead the substitution rate at fourfold degenerate sites with the substitution rate in pseudogenes. On the other hand, the rate of nucleotide substitution at nondegenerate sites is usually quite similar to the nonsynonymous rate. This is because the rate of transversional (nonsynonymous) substitution at twofold degenerate sites is on the average quite similar to the rates of transitional and transversional substitution at nondegenerate sites and also because in an average gene the number of twofold degenerate sites is about only one-fourth of the number of nondegenerate sites, so that a different rate at these sites would not affect the (weighted) average much. Therefore, there is no need to make a distinction between the substitution rate at nondegenerate sites and the nonsynonymous rate.

In computing the rates presented in table 2, we have assumed that the divergence time between two genes is the same as the divergence time between the two species from which the genes were obtained. In those genes that belong to multigene families there is a possibility that the genes compared were actually derived from a duplication before the species divergence or, in other words, that we compared paralogous rather than orthologous genes. This possibility exists for the histone genes, immunoglobulin genes, and the interferon-α and -β genes. Another possibility that can inflate the estimated rates is gene conversion between nonallelic genes of a multigene family.

Discussion
Estimation Methods

In our method, correction for multiple substitutions at the same site is made by using Kimura's (1980) two-parameter method, which takes into account only the possibility of unequal rates of transitional and transversional substitutions. Although this is usually the largest deviation from random substitution, other possibilities of nonrandom substitution may also be important. Our method will therefore tend to give underestimates when the degree of sequence divergence becomes large. In fact, this might be the reason why the rates of synonymous substitution in histones H4 and H3 obtained from comparisons between mammals
and chickens are considerably lower than those obtained from comparisons among mammals (table 2). Although it is not difficult to incorporate into our model more possibilities of nonrandom substitution, the computational procedure becomes complicated and the results are difficult to interpret. Moreover, incorporation of many parameters in a model tends to reduce its applicability when the degree of sequence divergence becomes large (Gojobori et al. 1982a; Tajima and Nei 1984). Of course, these are common difficulties in the estimation of the number of nucleotide substitutions (Gojobori et al. 1982a; Li et al. 1985).

We should also mention that our formula for the variances of \( A_i \) and \( B_i \) are approximations. We have neglected the covariance between \( P_i \) and \( Q_i \), because we assumed that \( P_i \) and \( Q_i \) are independent binomial variables. The covariance is, however, usually small compared with the variances of \( P_i \) and \( Q_i \). We also note that the classification of a site as synonymous or not depends on the sites adjacent to it. Therefore, changes in adjacent sites can affect the status of a site. This effect would tend to increase the variances of \( A_i \) and \( B_i \). The effect is probably negligible when \( A_i \) and \( B_i \) are small but may not be so when \( A_i \) and \( B_i \) are large.

We now compare our method with those of Miyata and Yasunaga (1980) and Perler et al. (1980). In our view, the latter two methods have three drawbacks. First, both methods assume that nucleotide substitution occurs randomly. This assumption tends to underestimate the number of substitutions, particularly if the degree of sequence divergence is large (Takahata and Kimura 1981; Gojobori et al. 1982a; Gojobori 1983). By contrast, our method allows for the difference between transitional and transversional rates, which is usually the largest deviation from random substitution (Li et al. 1984). Second, correction for multiple substitutions at the same site has not been done satisfactorily in previous methods. Perler et al.’s formulas are equivalent to assuming that, at twofold degenerate sites, substitutions can occur only between two states when considering synonymous substitutions and only among three states when considering nonsynonymous substitutions. For example, TTT is allowed to change to TTC, but not to TTA or TTG, in the estimation of synonymous rates. This assumption is incorrect. In Miyata and Yasunaga’s method, synonymous substitutions at twofold and fourfold degenerate sites are considered together and so are nonsynonymous substitutions at twofold degenerate and nondegenerate sites. They then use Jukes and Cantor’s (1969) method to make corrections for multiple substitutions. This is rather ad hoc. In our method, twofold degenerate sites, fourfold degenerate sites, and nondegenerate sites are considered separately, so that corrections for multiple substitutions can be done more rigorously. Moreover, in our method the error variances can also be derived in a more rigorous manner. (Actually, using \( P_i \), \( Q_i \), and \( L_i \), we have also included in our computer program a computational procedure similar to that of Miyata and Yasunaga. Of course, our weights for the possible paths between codons are different from theirs.) Third, Perler et al.’s assumption of equal probability for different pathways tends to underestimate the rate of synonymous substitution, particularly when the degree of sequence divergence becomes large. Miyata and Yasunaga do assign different weights for different pathways, but, as mentioned above, their weights are more in favor of nonsynonymous than synonymous substitutions. Therefore, their method also tends to underestimate the rate of synonymous substitution. As noted above, our weights are based on direct comparisons of DNA sequences and should therefore be more reasonable than those of Miyata and
Yasunaga. We should, however, emphasize that these three drawbacks would become serious only when the degree of sequence divergence becomes relatively large. In fairly closely related sequences—say, $K_S < 0.5$—their methods and ours are expected to give similar results.

In addition to the above, our method has two more advantages. First, it is easier to understand. The computational procedures in the other two methods are not easy to understand, and the classification of sites in Perler et al.'s method is quite complicated. Second, it provides more information on the substitution rates at sites with different degrees of coding degeneracy. This is useful for understanding the mechanisms of DNA evolution (see below).

The three methods have been applied to $\beta$-globin genes, and the results are shown in table 3. All three methods give very similar estimates for the number ($K_A$) of nonsynonymous substitutions per nonsynonymous site. This is to be expected, because the $K_A$ values are only 0.5 or smaller, so that the chance for multiple substitutions to occur at a site is small—or, in other words, the corrected values are close to the uncorrected values regardless of the method used. In the comparisons between mammals and in those between mammals and birds, the three methods also give similar estimates for the number ($K_S$) of synonymous substitutions per synonymous site. This is also not surprising, for the $K_S$ values are only about 0.8 or smaller. In the comparisons between mammals and frog, however, the three methods give very different estimates. As expected, the estimate by Perler et al.'s method is the smallest, while that by our method is the largest. Of course, one cannot conclude from these examples that our method performs better than the other two methods. To compare the relative performances, one will need to conduct a simulation study.

### Table 3

Numbers ($K_A$) of Nonsynonymous Substitutions per Nonsynonymous Site and Numbers ($K_S$) of Synonymous Substitutions per Synonymous Site between $\beta$-Globin Sequences

<table>
<thead>
<tr>
<th>Comparisons</th>
<th>Methods</th>
<th>$K_A$</th>
<th>$K_S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between mammal species</td>
<td>1</td>
<td>0.121</td>
<td>0.489</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.123</td>
<td>0.435</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.121</td>
<td>0.443</td>
</tr>
<tr>
<td>Mammals vs. birds</td>
<td>1</td>
<td>0.247</td>
<td>0.777</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.238</td>
<td>0.831</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.245</td>
<td>0.843</td>
</tr>
<tr>
<td>Mammals vs. frog</td>
<td>1</td>
<td>0.502</td>
<td>1.359</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.483</td>
<td>1.822</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.478</td>
<td>2.233</td>
</tr>
</tbody>
</table>

**Note.**—In this case the numbers of synonymous substitutions are obtained by taking the average of $P'$s and $Q$'s of the four comparisons before using eqs. (1) and (3) for correction.

*a* Perler et al. (1980), 1; Miyata and Yasunaga (1980), 2; and present authors, 3.

*b* Pairwise comparisons among human (Lawn et al. 1980), mouse (Konkel et al. 1979), rabbit (Hardison et al. 1979), and goat (Schon et al. 1981).


*d* Mammals (man, mouse, rat, and goat) and *Xenopus* (Williams et al. 1980).
Estimating ancient divergence is inherently unreliable, and most known factors tend to result in underestimation. For this reason, Perler et al.‘s conclusion that, in the evolution of globin and insulin genes, synonymous substitution was much slower in the period before mammalian divergence than in the more recent period has been challenged by Brown et al. (1982), Gojobori (1983), and Kimura (1983). The results in table 3 suggest that Perler et al.‘s method can indeed give a serious underestimate of Ks when comparing distantly related genes. Therefore, their conclusion is not warranted.

All the methods discussed here have to classify nucleotide sites in some way. These methods would be unreliable if the number of sites in each class (L0, L2, and L4 in the present model) changed considerably from time to time and from species to species. Fortunately, this does not appear to be the case. In our study, the number of sites for the same gene usually remains fairly stable across the species examined, even for distantly related ones.

Relative Likelihoods of Codon Changes

Many previous studies have been devoted to the subject of amino acid exchangeability, primarily in an attempt to understand the biochemical roles of different amino acids in various proteins (e.g., Zuckerkandl and Pauling 1965; Epstein 1967; McLachlan 1971). As a consequence, many amino acid distance matrices have been proposed to correlate the evolutionary exchangeability with known physicochemical differences between amino acids (Sneath 1966; McLachlan 1971; Grantham 1974; Miyata et al. 1979). All these studies have been based on amino acid sequence data and hence have two limitations: (1) It is difficult to formulate accurately the expected number of amino acid changes without knowing the underlying codons. Previous studies assumed that the expected number of changes between aa and aa, is proportional to fαfα, where fα is the frequency of aa. This is obviously only an approximation. (2) Amino acid sequence data per se provide no information on synonymous changes. Our analysis of DNA sequence data was motivated to overcome these two limitations and, in our opinion, provides a more reliable estimate of codon or amino acid exchangeability.

From table 1, we see a strong correlation between relative likelihood of amino acid interchanges and amino acid distances on the part of nonsynonymous substitution. This suggests that Grantham’s (1974) indices are quite adequate in predicting amino acid exchangeability. It is also clear that purifying (negative) selection arising from functional constraints is a major force at this level of evolution. On the other hand, the relative likelihood of synonymous changes is much greater than the value extrapolated from amino acid changes in protein sequences. For this reason, Miyata and Yasunaga’s (1980) method would give a pathway with synonymous substitutions a lower weight than expected.

The results of table 1 are obtained by taking the average of all minimum distance solutions, when there is more than one solution. We note that all minimum distance solutions require the same smallest number of substitutions, regardless of the nature of codon changes involved. In reality, evolution is more likely to have taken a pathway that requires more synonymous codon changes and fewer radical amino acid replacements. Taking the average means that we may still underestimate the likelihood of synonymous (and perhaps even conservative) changes and overestimate that of the radical changes. The advantage of using many sequences and
inferring their ancestral codons is that such ambiguous cases can be reduced. It also enables us to formulate the expected number of codon changes more precisely.

Finally, the results of table 1 also have a bearing on parsimonious phylogenetic reconstruction using amino acid or DNA sequence data (e.g., Moore et al. 1973; Czeluźniak et al. 1982). In using the maximum parsimony method, authors usually make no distinction between the types of codon changes inferred. We point out that considering the relative likelihoods of different codon changes can increase the discriminative power of the maximum parsimony principle. To see this point, let us consider a simple hypothetical example in which the codons of four DNA sequences at a particular homologous codon position are, respectively, CCT (Pro), CCG (Pro), CGT (Arg), and CGG (Arg). Designate the four sequences as a, b, c, and d. If all nucleotide substitutions are given the same weight, the two phylogenetic trees \([\{(a, b), (c, d)\}\) (which means that a joins b, c joins d, and then their ancestors join each other) and \([\{(a, c), (b, d)\}\) are both most parsimonious, each requiring three nucleotide substitutions. However, the first tree requires two synonymous substitutions and only one nonsynonymous substitution (Pro for Arg or Arg for Pro), whereas the second tree requires one synonymous substitution and two nonsynonymous substitutions (both are replacements of Arg by Pro or replacements of Pro by Arg). Since the likelihood of a synonymous substitution is nearly three times that of the replacement of Pro by Arg or Arg by Pro (a moderately conservative replacement), the first tree may be considered more parsimonious if greater weights (distance values) are given to nonsynonymous substitutions.

To illustrate how to apply the “weighted” parsimony method to real data, let us use the α-globin genes from man (h), mouse (m), rabbit (r), and goat (g, and g2 representing alleles 1 and 2) as an example. Consider the trees \([\{(h, m), (g1, g2)\}\), \([\{(h, r), (m, h)\}, (g1, g2)\})\ and \([\{(m, r), (h, g1, g2)\})\). At the nucleotide level, the first two trees require 159 substitutions and the third 160 substitutions. At the codon level, the first tree requires 95.6 synonymous changes, 21.3 conservative, 32.1 moderately conservative, 10 moderately radical, and 0 radical amino acid replacements. The numbers for the second tree are 91.3, 19.9, 35.7, 11.6, and 0.5, respectively, and those for the third tree are 93.8, 19.7, 34.9, 11.0 and 0.5, respectively. Since a conservative amino acid change is equivalent to \(2.235/0.853 = 2.62\) synonymous changes, a moderately conservative one is equivalent to \(2.235/0.767 = 2.91\) synonymous changes, and so on, we may convert all the nucleotide substitutions into the equivalent of synonymous ones. The first tree, therefore, requires the equivalent of 323.4 synonymous substitutions, the second tree requires 351.0, and the third tree requires 345.8. One may therefore choose the first tree over the other two on the ground of codon changes. This choice is, of course, quite tentative, for only one gene is used.

Mechanisms of DNA Evolution

The results in table 2 are, in our opinion, quite compatible with the neutral mutation–random drift hypothesis (Kimura 1968; King and Jukes 1969). This hypothesis predicts that the rate of evolution will be higher for molecules or parts of a molecule with weak or no selective constraints and low for those that are tightly constrained. Among the coding regions of a gene, nondegenerate sites are expected to be most tightly constrained because all mutations at these sites are nonsynonymous, so that they are subject to constraints at both the RNA and
protein levels. The constraints at the protein level would be particularly strong, for natural selection acts on phenotypes of the organism for which the protein structure and function play a decisive role. Thus, the substitution rate is expected to be lowest at nondegenerate sites. This is indeed the case (table 2). The rate should be higher at twofold degenerate sites, for a fraction of the mutations at these sites are synonymous, so that these sites should be less stringently constrained than nondegenerate sites. The rate should be even higher at fourfold degenerate sites, for all mutations at these sites are synonymous. These two predictions are also seen to be true.

Under the neutralist view, we can also give a plausible explanation for why the rate of nonsynonymous substitution varies greatly among different kinds of genes. For those genes, such as histone and actin genes, that have extremely low nonsynonymous rates (table 2), we may assume that they are subject to tight selective constraints because their functions are important and require very specific protein structures. According to this view, nearly every nonsynonymous mutation in histone and actin genes is deleterious and would not become fixed in the population. On the other hand, for the genes, such as those coding for relaxin, interferons, and immunoglobulins, that have high nonsynonymous rates (table 2), we may assume that they are subject to fewer or weaker selective constraints because their functions do not require very specific protein structures or because their functions are not very important or can be carried out by some other genes, i.e., these genes are "dispensable" (Wilson et al. 1977). We also noted above that the variation in synonymous rate among genes is large, though much smaller than that in nonsynonymous rate. Under the neutralist view, this variation can be explained by assuming that it is largely due to chance effects and partly due to variation in selective constraints among genes. It now appears that synonymous mutations are not completely free of selective constraints but may be constrained by such factors as tRNA availability (Ikemura and Ozeki 1983) and the binding strength between codon and anticodon (Grosjean and Fiers 1982).

Although some of the above arguments were based on assumptions that remain to be substantiated by experimental data, the neutralist view can apparently give a consistent explanation for the various features of DNA evolution. In contrast, the selectionist view cannot give a consistent explanation. To explain the negative correlation between rate and selective constraints by this view, one would have to assume that the probability of a mutation being selectively advantageous is higher in functionally less important DNA regions. It is, however, difficult to imagine how the function of a gene can be constantly improved by synonymous mutations. One may concede that neutral or nearly neutral mutations are predominant at synonymous sites while concurrently maintaining that the majority of amino acid substitutions in protein evolution have been due to advantageous mutations (see, e.g., Goodman 1982). But it is again difficult to imagine why advantageous mutations should occur more often in less important coding regions, such as those coding for the signal peptides and the insulin C and relaxin C peptides—or, in other words, why these latter regions have constantly been improved by new mutations. Furthermore, it is difficult to imagine why the rate of improvement in function should be faster in a less important gene than in a more important one. In short, the selectionist view cannot give a satisfactory explanation for the negative correlation between rate and functional importance.
Table 4
Average Substitution Rates in Pseudogenes and at Fourfold Degenerate, Twofold Degenerate, and Nondegenerate Sites of Functional Genes

<table>
<thead>
<tr>
<th>Parameter of Comparison</th>
<th>Pseudogenes</th>
<th>Fourfold Degenerate Sites</th>
<th>Twofold Degenerate sites</th>
<th>Nondegenerate Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate ( \times 10^9 )</td>
<td>4.85</td>
<td>4.18</td>
<td>2.26</td>
<td>0.94</td>
</tr>
<tr>
<td>Ratio to the pseudogene rate</td>
<td>1.00</td>
<td>0.86</td>
<td>0.47</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Under the neutralist view, one can use the substitution rate in pseudogenes as the neutrality standard to infer the stringency of selective constraints in various gene regions. From the studies of Li et al. (1981), Li (1983), and Gojobori and Nei (1984), we obtain an average rate of \( 4.85 \times 10^{-9} \) substitutions per nucleotide site per year. The ratios of the average substitution rates at fourfold degenerate sites, twofold degenerate sites, and nondegenerate sites to the pseudogene rate are shown in Table 4. Fourfold degenerate sites are on the average subject to only very weak constraints, because their substitution rate is only 14% lower than the pseudogene rate. Twofold degenerate sites are subject to certain constraints, for they evolve only half as fast as pseudogenes. Nondegenerate sites are generally tightly constrained, for their average substitution rate is only one-fifth of the pseudogene rate. If we assume that the nucleotide substitutions at nondegenerate sites are totally due to neutral mutations, then the proportion of neutral mutations at nondegenerate sites is approximately 20%. This is quite similar to Kimura's (1983b) estimate of 0.14 ± 0.06. Of course, the proportion would vary from gene to gene, ranging from 62% in interferon-\( \gamma \) to almost zero in histone H4 (Table 2). Similarly, we can infer the stringency of selective constraints in various genes. For example, albumin has previously been used as an example of rapid evolution at the amino acid level, and the argument for this is that albumin is a dispensable protein (Wilson et al. 1977). Although the nonsynonymous rate in the gene for albumin is not low, being \( 0.92 \times 10^{-9} \) (Table 2), it is only 19% of the pseudogene rate. Thus, it appears that albumin is subject to fairly strong selective constraints. Among the genes listed in Table 2, interferon-\( \gamma \) has the highest nonsynonymous rate, \( 2.80 \times 10^{-9} \). This is 58% of the pseudogene rate. Thus, it appears that, even in a rapidly evolving gene, a substantial fraction of nonsynonymous mutations are deleterious.

Acknowledgments

This work was supported by research grants from the National Institutes of Health and the National Science Foundation. We thank W. M. Fitch for suggestions.

LITERATURE CITED


WALTER M. FITCH, reviewing editor

Received July 26, 1984; revised October 29, 1984.
The Effects of Mispair and Nonpair Correction in Hybrid DNA on Base Ratios (G + C Content) and Total Amounts of DNA

B. C. Lamb
Department of Pure and Applied Biology, Imperial College, London

Base ratios and total DNA amounts can vary substantially between and within higher taxa and genera, and even within species. Gene conversion is one of several mechanisms that could cause such changes. For base substitutions, disparity in conversion direction is accompanied by an equivalent disparity in base ratio at the heterozygous site. Disparity in the direction of gene conversion at meiosis is common and can be extreme. For transitions (which give purine [R]/pyrimidine [Y] mispairs) and for transversions giving unlike R/R and Y/Y mispairs in hybrid DNA, this disparity could give slow but systematic changes in G + C percentage. For transversions giving like R/R and Y/Y mispairs, it could change AT/TA and CG/GC ratios. From the extent of correction direction disparity, one can deduce properties of repair enzymes, such as the ability (1) to excise preferentially the purine from one mispair and the pyrimidine from the other for two different R/Y mispairs from a single heterozygous site and (2) to excise one base preferentially from unlike R/R or Y/Y mispairs. Frame-shifts usually show strong disparity in conversion direction, with preferential cutting of the nonlooped or the looped-out strand of the nonpair in heterozygous h-DNA. The opposite directions of disparity for frame-shifts and their intragenic suppressors in Ascbolus suggest that repair enzymes have a strong, systematic bias as to which strand is cut. The conversion spectra of mutations induced with different mutagens suggest that the nonlooped strand is preferentially cut, so that base additions generally convert to mutant and deletions generally convert to wild-type forms. Especially in nonfunctional or noncoding DNA, this could cause a general increase in DNA amounts. Conversion disparity, selection, mutation, and other processes interact, affecting rates of change in base ratios and total DNA.

Introduction

It has long been established (e.g., Sueoka 1961) that organisms may have quite different base ratios in DNA, as measured by molar percentage of guanine plus cytosine (G + C). The G + C percentage also differs between different molecules within an organism and between different parts of the same molecule. Table 1 shows the wide range of G + C content in different groups. Table 2 shows the large variation in G + C percentage even within a genus (e.g., from 28% to 74% for the bacterium Sarcina, from 30% to 60% for the fungus Candida, and from 19% to

1. Key words: gene conversion, base ratios, DNA amounts, repair of hybrid DNA. Abbreviations: h-DNA = hybrid DNA, Y = pyrimidine, R = purine, m = mutant, + = wild-type.

Address for correspondence and reprints: Dr. B. C. Lamb, Department of Pure and Applied Biology, East Bihi, Imperial College, London SW7 2BB, England.

© 1985 by The University of Chicago. All rights reserved.
0737-4038/85/0202-001252.00

175
Table 1
Variation in Base Ratios within Groups

<table>
<thead>
<tr>
<th>GROUP</th>
<th>G + C PERCENTAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>Bacterial phages</td>
<td>27.7</td>
</tr>
<tr>
<td>Viruses of chordates</td>
<td>34.8</td>
</tr>
<tr>
<td>Bacteria</td>
<td>22.8</td>
</tr>
<tr>
<td>Fungi</td>
<td>22.0</td>
</tr>
<tr>
<td>Algae</td>
<td>36.9</td>
</tr>
<tr>
<td>Bryophytes</td>
<td>41.2</td>
</tr>
<tr>
<td>Gymnospermae</td>
<td>34.8</td>
</tr>
<tr>
<td>Angiosperms</td>
<td>34.0</td>
</tr>
<tr>
<td>Protozoa</td>
<td>19.0</td>
</tr>
<tr>
<td>Porifera</td>
<td>34.8</td>
</tr>
<tr>
<td>Coelenterates and Aschelminthes</td>
<td>36.8</td>
</tr>
<tr>
<td>Mollusca</td>
<td>30.7</td>
</tr>
<tr>
<td>Arthropoda</td>
<td>31.9</td>
</tr>
<tr>
<td>Echinodermata</td>
<td>35.0</td>
</tr>
<tr>
<td>Chordata</td>
<td>36.2</td>
</tr>
</tbody>
</table>

Source.—Compiled from data quoted by Normore (1976).

32% for the protozoan *Tetrahymena*). Even within a species, wide variations in G + C percentage may occur between different wild-type isolates (e.g., from 29% to 51% for *Lactobacillus delbrueckii*) or between wild-type and mutant strains (e.g., from 30% to 71% for *Micrococcus aureus*). Base ratios are clearly capable of evolutionary change.

Established processes affecting base ratios and/or total DNA amounts include mutation (base substitutions; additions or deletions of from one to many bases), selection, genetic drift, unequal crossing-over, duplication, amplification, and karyotype changes (e.g., gain or loss of whole chromosomes or sets of chromosomes). The amount of change depends on their frequency and nature (e.g., length of a deletion or the G + C percentage of an amplified region in relation to total G + C percentage). In the present study, it is proposed that disparity in the direction of gene conversion is another such process, capable of changing base ratios and total DNA amounts both in discrete sequences and systematically throughout the genome, including noncoding regions.

Gene conversion involves the formation of hybrid (heteroduplex) DNA, with varying frequencies of enzymatic correction of mispairs from heterozygous base substitutions or of nonpairs from heterozygous frame-shifts. In eukaryotes, gene conversion occurs at meiosis with frequencies varying from less than 1 in 10^3 to >30% at a site (Lamb and Helmi 1982; Lamb 1984). Conversion-like processes and hybrid DNA formation are also known in prokaryotes (for reviews, see Catcheside [1977]; Stahl [1979]; Whitehouse [1982]). Meiotic conversion will be the main consideration here, although the much lower frequencies of mitotic gene conversion could have effects, as could processes involving h-DNA in prokaryotes.

To be effective, conversion requires polymorphism and heterozygosity, which are widespread in most eukaryotes (e.g., Nevo 1978). The effects of conversion on
Table 2
Variation in Base Ratios within Genera and within Species

<table>
<thead>
<tr>
<th>TAXON</th>
<th>G + C PERCENTAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>Bacterial genera:</td>
<td></td>
</tr>
<tr>
<td><em>Achromobacter</em></td>
<td>39.5</td>
</tr>
<tr>
<td><em>Flavobacterium</em></td>
<td>30.0</td>
</tr>
<tr>
<td><em>Bacillus</em></td>
<td>31.7</td>
</tr>
<tr>
<td><em>Clostridium</em></td>
<td>25.3</td>
</tr>
<tr>
<td><em>Cytophaga</em></td>
<td>33.0</td>
</tr>
<tr>
<td><em>Micrococcus</em></td>
<td>28.5</td>
</tr>
<tr>
<td><em>Sarcina</em></td>
<td>28.0</td>
</tr>
<tr>
<td><em>Mycoplasma</em></td>
<td>22.8</td>
</tr>
<tr>
<td><em>Pseudomonas</em></td>
<td>36.1</td>
</tr>
<tr>
<td><em>Spirillum</em></td>
<td>28.0</td>
</tr>
<tr>
<td><em>Vibrio</em></td>
<td>29.5</td>
</tr>
<tr>
<td>Bacterial species:</td>
<td></td>
</tr>
<tr>
<td><em>Bacillus subtilis</em></td>
<td>41.5</td>
</tr>
<tr>
<td><em>Moraxella bovis</em></td>
<td>41.0</td>
</tr>
<tr>
<td><em>Bacterium paracoli</em></td>
<td>48.0</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>49.8</td>
</tr>
<tr>
<td><em>Lactobacillus delbrueckii</em></td>
<td>28.9</td>
</tr>
<tr>
<td><em>Micrococcus aureus</em></td>
<td>30.2</td>
</tr>
<tr>
<td><em>Sarcina afermentans</em></td>
<td>32.0</td>
</tr>
<tr>
<td>Fungi:</td>
<td></td>
</tr>
<tr>
<td><em>Achlya</em></td>
<td>44.5</td>
</tr>
<tr>
<td><em>Aspergillus</em></td>
<td>46.0</td>
</tr>
<tr>
<td><em>Candida</em></td>
<td>30.0</td>
</tr>
<tr>
<td><em>Mucor</em></td>
<td>29.5</td>
</tr>
<tr>
<td><em>Penicillium</em></td>
<td>47.0</td>
</tr>
<tr>
<td><em>Rhodotorula</em></td>
<td>47.5</td>
</tr>
<tr>
<td><em>Saccharomyces</em></td>
<td>30.5</td>
</tr>
<tr>
<td>Algae:</td>
<td></td>
</tr>
<tr>
<td><em>Polytoma</em></td>
<td>42.0</td>
</tr>
<tr>
<td>Protozoa:</td>
<td></td>
</tr>
<tr>
<td><em>Tetrahymena</em></td>
<td>19.0</td>
</tr>
</tbody>
</table>

SOURCE.—Compiled from data quoted by Normore (1976).
NOTE.—Numbers in parentheses are percentages for mutant strains.
*Now called Staphylococcus aureus.*

base ratios and DNA amounts will be directly related to conversion frequencies and to the extents of disparity, parameters surveyed by Lamb and Helmi (1982) and Lamb (1984). Disparity in conversion direction is generally very common and often large, ranging from nearly all conversions being from mutant (m) to wild type (+), to nearly all conversions being from + to m; sites with no disparity also occur.

The relevant parameters for meiotic conversion were defined by Lamb and Helmi (1982). Let \( c \) be the conversion frequency, that is, the fraction of meioses with aberrant segregation ratios (e.g., +:m or A:a) from gene conversion at the site considered. Let \( b \) be the frequency of a particular allele (+ for +/m and A for A/a) in the products of meiotic tetrads or octads with aberrant segregation ratios from
conversion (ratios other than 2+:2m or 4+:4m). Let d be the disparity in direction of conversion, measured as \( b = -0.5 \). In fungi, Lamb (1984) found that c ranged from 0.0004 to 0.302; that \( b \) ranged from 0.24 to 0.83, and that \( d \) ranged from -0.26 to +0.33. For \( d = 0 \) there is no disparity, while \( d = -0.25 \) is the most extreme value normally expected from disparity favoring m, with all conversion to m; \( d = +0.25 \) is the most extreme value normally expected from disparity favoring +, with all conversion to +, although slightly more extreme values occasionally occur (Lamb and Helmi 1982). Transfection experiments in prokaryotes have shown strong preferential excision of particular strands in mismatch repair, sometimes related to and sometimes independent of the degree of methylation of the DNA (e.g., Pukkila et al. 1983; Dohet et al. 1984). Transfection experiments with Bacillus and Ustilago (reviewed in Whitehouse [1982; pp. 331–332]) suggest that the nature of the mismatch and other features of the DNA, such as neighboring base sequences, influence which base is excised during correction.

Chemical studies were made of mispairs and nonpairs in short, defined heteroduplexes, e.g., by Dodgson and Wells (1977), Cornelis et al. (1979), Patel et al. (1982). In h-DNA, at least some mispairs have bases within the helix, with intact adjacent base pairs—although with local perturbations in base stacking with neighboring bases. Even the extra base of an addition frame-shift may stack within the helix, but for heterozygous frame-shifts here the strand with the extra base (+ for deletions, m for additions) will for convenience be termed the "looped-out strand" as opposed to the "nonlooped strand" (see fig. 1).

**Two Causes of Disparity in the Direction of Conversion and Their Relative Importance**

Disparity in the direction of conversion (e.g., unequal proportions of + and m products from conversion in a +/m heterozygote) could arise either during hybrid DNA correction, because a repair enzyme preferentially excised one strand from h-DNA (e.g., giving more conversion to m than to + if the + strand is preferentially excised), or during h-DNA formation if asymmetric h-DNA is formed and is formed preferentially on one type of chromatid (e.g., preferentially on the + rather than the m chromatid). The relative frequencies of symmetric h-DNA (giving +/m on

<table>
<thead>
<tr>
<th>Specimen sequence of + gene</th>
<th>Point deletion from +</th>
<th>Point addition to +</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>AT</td>
<td>ACCT</td>
</tr>
<tr>
<td>TGA</td>
<td>TA</td>
<td>TUGA</td>
</tr>
</tbody>
</table>

Hybrid DNA:

```
1 C      +
4 TA  m (deletion)
```

```
1 ACT    +
4 TGA  m (addition)
```

```
3 AT  m (deletion)
2 TA  +
```

```
3 ACT  m (addition)
2 TUGA +
```

**FIG. 1.**—Nonpairs for heterozygous frame-shifts in hybrid-DNA. Numbers 1, 2, 3, and 4 represent the individual DNA strands of the two involved homologous, nonsister chromatids. In symmetric h-DNA for +/m, both nonpairs form, one in each involved chromatid; in asymmetric h-DNA, only one of the two nonpairs forms at a time.
two chromatids, one originally + and one originally m in this example) and of asymmetric h-DNA (giving h-DNA in only one chromatid at the point of heterozygosity considered) vary greatly in different studies (see Whitehouse 1982, pp. 298–312). Disparity in the direction of conversion will not arise during h-DNA formation if all h-DNA is symmetric or if asymmetric h-DNA has no chromatid preference, but it could arise during correction, however the h-DNA is formed.

In theory, conversion disparity from asymmetric h-DNA chromatid preference is likely to be rare. If the known +/m difference is the only point of heterozygosity in the chromosome region, then there is no reason for asymmetric h-DNA chromatid preference unless the heterozygosity happens to be in some DNA sequence specifically concerned with asymmetric h-DNA initiation. If there are other nearby points of heterozygosity, either known or cryptic (see Lamb 1975), they should also have no effect on asymmetric h-DNA chromatid preference unless in such a special sequence. As discussed in the following section, there is strong evidence from *Ascobolus* frameshifts that conversion disparity from asymmetric h-DNA chromatid preference is rare. It is therefore likely that most, perhaps almost all, disparity in the direction of conversion arises from disparity in the direction of correction, permitting conclusions about repair enzyme properties to be drawn from data on the direction of conversion. The main analyses therefore exclude asymmetric h-DNA chromatid preference, which is discussed later, but they include both asymmetric h-DNA with no chromatid preference and symmetric h-DNA.

**Frame-Shifts: Disparity Data and Effects of Repair on Total DNA Content**

Nonpairs produced by heterozygous addition or deletion frame-shifts in +/m h-DNA are shown in figure 1. In symmetric h-DNA, both nonpairs form, one in each involved chromatid. In asymmetric h-DNA, only one of the two nonpairs forms at a time, but the two types of nonpair should be equally frequent over a population of meioses.

If the repair system for nonpairs showed no preference between cutting the looped-out or the nonlooped strand, there would be no disparity in conversion direction for frame-shifts, but that cannot be a general property of repair enzymes because frame-shifts generally have strong disparity, as shown in table 3 and by Lamb (1984). Indeed, frame-shifts usually have significantly more disparity than do base substitutions, as shown in table 3, where average absolute disparities for frame-shifts range from 0.180 to 0.235 for five loci, very high in relation to the expected maximum absolute disparity of ~0.25. While most frame-shifts show strong disparity, a few have no significant disparity.

This strong disparity could result from three possible repair enzyme properties: (1) a general tendency to excise the looped-out strand rather than the nonlooped one, (2) a tendency to excise the nonlooped strand, and (3) a tendency to excise either the looped-out strand or the nonlooped strand, depending on the site. It is difficult to see how (3) could arise if the nature of the base looped out had no effect on excision, and it cannot be explained by the effects of neighboring base sequences on which type of strand is cut. For example, in figure 1, suppose that strands 1 and 3 (which have the same polarity and the same base sequence except at the point of mutation) had some neighboring base sequence causing those strands to be preferentially cut during repair. For the point deletion, cutting strand 1 cuts the looped-out (+) strand and cutting strand 3 cuts the nonlooped (m) strand in the other
Table 3
Disparity in Gene Conversion Direction (d): Comparison of Frame-shift and Base Substitution Mutations

| LOCUS                  | No. of Mutations | Mean Absolute Value of d | \( \sigma \) of | Limits of \( |d| \), (Actual Values) | No. of Mutations | Mean Absolute Value of d | \( \sigma \) of | Limits of \( |d| \), (Actual Values) |
|------------------------|------------------|--------------------------|----------------|-----------------------------|------------------|--------------------------|----------------|-----------------------------|
|                        |                  |                          | \( |d| \)      | Greatest                  | Least | Mean d            |                  | \( |d| \)      | Greatest                  | Least | Mean d            |
| *Sordaria brevicollis* (Yu-Sun et al. 1977): |                  |                          |               |                            |       |                   |                  |                          |             |                   |             |                   |
| grey-3                 | 7                | .235***                  | .014          | -.252                     | -.207 | -.235             | 4               | .095          | .026                      | -.125 | +.068             | +.032 |
| grey-4                 | 4                | .180**                   | .020          | -.202                     | -.163 | -.180             | 6               | .066          | .065                      | +.175 | +.009             | +.052 |
| grey-5                 | 11               | .227**                   | .033          | -.263                     | -.147 | -.143             | 7               | .029          | .023                      | -.062 | -.004             | -.007 |
| *Ascomobolus immersus* (Leblon 1972a): |                  |                          |               |                            |       |                   |                  |                          |             |                   |             |                   |
| b1                     | 10               | .220***                  | .040          | -.262                     | +.154 | -.107             | 21              | .064          | .040                      | +.129 | +.013             | +.037 |
| b2                     | 24               | .192***                  | .070          | -.249                     | -.030 | -.164             | 7               | .055          | .021                      | +.083 | +.032             | +.045 |

Note.—Presumed frame-shifts have a very low proportion of conversion asci with postmeiotic segregation (5:3, 3:5) compared with conversion asci with meiotic segregation (6:2, 2:6), whereas the proportion of postmeiotic segregations is not nearly so low for presumed base substitutions.

*, **, *** Differences between frame-shifts and base substitutions for mean absolute \( d \) significant at \( P = 5\% \), 1\%, and 0.1\%, respectively, using a "t" test.
chromatid, so that the effects of neighboring base sequences cancel each other out in the two types of nonpair and do not cause overall conversion disparity.

There is evidence from *Ascobolus* and *Sordaria* that repair systems do preferentially cut one particular type of strand in nonpairs. A frame-shift mutation and a nearby intragenic frame-shift suppressor of that mutation (where the suppressor on its own is also mutant) must be of opposite sign, i.e., one must be a deletion and the other an addition. If the repair system generally cuts one type of strand, then a frame-shift mutation should generally have the opposite direction of conversion disparity as its intragenic frame-shift suppressors, and different intragenic frame-shift suppressors of the same frame-shift mutation should generally have the same direction of disparity as each other, since the direction of disparity will depend just on whether they are additions or deletions.

In *A. immersus*, frame-shift *w1-3C1* had a disparity of +0.17, whereas its intragenic suppressor, *w1-WSu*, had a *d* of −0.11 (Lamb 1984). The data of Rossignol et al. (1979) from the b2 locus of *A. immersus* are very conclusive; values from their table 8 have been changed here to correspond to *d*. Frame-shift E1, with a *d* of −0.22, had two intragenic suppressors, both with a *d* of ∼ +0.20. Frame-shift A0, with a *d* of −0.22, had four intragenic suppressors, with *d* values ranging from +0.15 to +0.23. Frame-shift F0, with a *d* of −0.10, had 13 intragenic suppressors, with *d* values ranging from +0.05 to +0.13. Thus, for 20 cases of intragenic suppressors of frame-shifts, every one had a direction of disparity opposite to that of the original mutation, and different frame-shift suppressors of the same mutation always had the same direction of disparity as each other. This is entirely consistent with the hypothesis that frame-shifts of one sign convert preferentially in one direction and that frame-shifts of the opposite sign correct preferentially in the other direction. This accords completely with either property (1) or (2) but not with (3), because with (3) some frame-shift suppressors could have the same direction of disparity as the mutations they suppress, and different suppressors of the same frame-shift could have opposite directions of disparity.

Evidence for distinguishing between (1) and (2) comes from the relation between a mutation's conversion properties and the mutagen used to induce it. For the *b1* and *b2* loci of *A. immersus*, Leblon (1972a) found that all 25 ICR-170-induced mutations had a disparity favoring the mutant. For *grey-3*, *grey-4* and *grey-5* in *Sordaria brevicollis*, Yu-Sun et al. (1977) found that of 21 ICR-170 mutations, 19 had a disparity favoring the mutant and two had disparity favoring +. At *b2* in *A. immersus*, Rossignol et al. (1979) found that all six ICR-170 mutations had a disparity favoring m and that 18 EMS mutations had a disparity favoring +. Lamb and Ghikas (1979) found less clear correlations. This general—but not total—consistency of disparity direction with mutational origin is good evidence that frame-shifts of one sign usually have the same preferential direction of correction in heterozygous h-DNA. Leblon (1972b) concluded that his ICR-170 mutations were single base additions (see also Yu-Sun et al. 1977), and the EMS mutations of Rossignol et al. (1979) are probably single base deletions (although other EMS mutations may be of other types). This implies that base additions preferentially convert to mutants and that deletions preferentially convert to wild types, in accordance with property (2). That is, for single base nonpairs in h-DNA, the correction system preferentially cuts the nonlooped strand, so that gene conversion disparity should usually increase the amount of DNA in a genome, especially in noncoding regions or other parts where selection may be weak.
While the data cited so far have been for single base additions and deletions, longer deletions, such as 300–400 nucleotides, can also convert and show disparity in either direction (+ or −) in yeast, and a deletion of the whole cyclic locus showed conversion (see Lamb 1984, Whitehouse 1982, pp. 283–284). Disparity in direction of conversion for deletions or additions several hundreds of nucleotides long could obviously lead to much more rapid changes in total DNA content, especially in noncoding DNA regions, than could single base changes.

A preference for cutting the nonlooped strand would affect total DNA but would not affect G + C percentage unless the nature of the base looped out affected correction. As shown in figure 1, the two nonpairs for a heterozygous frame-shift are different, with one having a purine looped out and one having a pyrimidine looped out. Any effects of the nature of the base on excision would reduce conversion disparity; for example, always excising the looped strand when a purine is looped out and always excising the nonlooped strand when a pyrimidine is looped out would give no +/− disparity in the usual case, where the two types of nonpair are equally frequent. The typically strong disparities for frame-shifts suggest that, if they occur, any effects due to the nature of the base that is looped out on excision are small.

Most cases of frame-shift disparity must arise from disparity in correction direction, not from chromatic preference in asymmetric h-DNA formation, for theoretical reasons given above and because the latter cause should give equal amounts of disparity for base substitutions and frame-shifts, which is not the case. Similarly, chromatic preference would not lead to any correlation between mutagen used and direction and amount of disparity, and it would certainly not result in the observed opposite directions of disparity for frame-shifts and their intragenic suppressors. Thus, although chromatic preference in asymmetric h-DNA formation is possible, it must be very much rarer than disparity in correction direction if it is to account for various observed conversion properties.

**Base Substitutions: Disparity Data and Repair Properties**

For heterozygous base substitutions, one has to consider the correction of mispairs rather than nonpairs. Suppose that at meiosis one chromatid has a + allele with an AT base pair at the point of heterozygosity and that the homologous nonsister chromatid with the − allele has CG there (the superscripts here numbering the four DNA strands, with 1 and 3 having the same polarity). Symmetric h-DNA, formed on the Holliday (1964) or Meselson and Radding (1975) types of model, would have mispair AG on one chromatid and mispair CT on the other. Asymmetric h-DNA would contain either AG or CT, depending on which strand of which chromatid invaded the nonsister chromatid. Over a population of meioses with asymmetric h-DNA at that site, AG and CT mispairs should be of approximately equal frequency. Repair properties have to be considered for three types of mispair. In +/− heterozygotes, all transitions give R/Y mispairs in h-DNA, e.g., AT in + and GC in − gives two different R/Y mispairs, AC and GT. Some transversions give unlike R/R and Y/Y mispairs, e.g., AT(+o) and CG(−), giving AG and CT mispairs. Other transversions give like R/R and Y/Y mispairs, e.g., AT(+) and TA(−), giving AA and TT mispairs.
Table 3 shows the extent of disparity for base substitutions at five loci in the fungi *Ascocholus* and *Sordaria* in terms of locus averages; individual sites often showed much more extreme values. Thus, in the *Sordaria* data of Yu-Sun et al. (1977), grey-3 RW9 had a $d$ of $-0.125$, while grey-4 YS10 had a $d$ of $+0.175$, compared with approximate most expected values of $-0.25$ and $+0.25$ (Lamb and Helmi 1982) and their locus averages for mean absolute disparity of 0.095 and 0.066, respectively. Even more extreme disparity values occurred for base substitutions in the *Ascocholus* data of Lamb and Ghikas (1979): BHo had a $d$ of $-0.246$ and Bw.4.3 had a $d$ of $+0.234$, showing that base substitutions can have extreme disparity. Some have no significant disparity, but many have disparities significant at the 5%, 1%, or 0.1% levels. Other data with similar trends were analyzed by Lamb (1984).

These data on disparity enable one to draw conclusions about repair enzymes. Like mispairs (e.g., AA) could not show correction disparity caused by the nature of the bases (since they are identical) but might show individual disparity caused by the effects of neighboring sequences. With equal numbers of the two types of mispair, however, these individual disparities would tend to cancel each other out.

For example, $A\hat{T}(+)$ and $T\hat{A}(m)$ form AA and TT mispairs, where strands 1 and 3 have the same polarity and base sequence except at the point of mutation. A tendency to excise $A$ preferentially from AA, giving correction to $T\hat{A}(m)$, should be counterbalanced by a tendency to excise T from TT, giving correction to $A\hat{T}(+)$. Like mispairs are therefore unlikely to give overall conversion direction disparity for $+/m$ and could account for some sites with no significant disparity. The frequency of disparity is so high that both remaining types of mispair—unlike R/R or Y/Y, and R/Y—must be capable of giving correction disparity.

At least some repair enzymes must therefore show a preference as to which base is excised from R/Y and unlike R/R and Y/Y mispairs. Because of equal numbers of the two types of R/Y mispairs in a cross, disparity in conversion direction cannot be produced by a repair enzyme having the same preference for a particular type of base (say, for excising purines) from all R/Y mispairs. For example, with preferred excision of purines and equal numbers of AC and GT mispairs, ACs would largely be corrected to GC($m$) and GTs would largely be corrected to AT($+$), giving no overall disparity. Disparity could be produced by a preferential excision of one type of base (say, purines) if this occurred to different extents for different R/Y mispairs. The extreme case would be complete preference for excising the purine from one mispair and no preference at the other, e.g., ACs giving all GC($m$) but GTs giving 50% GC($m$) and 50% AT($+$), giving an absolute disparity of $\sim 0.125$. The results quoted show that absolute disparities for base substitutions can sometimes far exceed 0.125; additionally, correction often fails to occur, particularly for base substitutions (see Lamb 1984), and repair failure reduces any disparity by giving one + and one $m$ product after replication of a $+/m$ mispair. The occurrence of absolute $d$ values $>0.125$, in spite of some repair failure, strongly suggests that some repair enzymes are able to excise the purine preferentially from one R/Y mispair and the pyrimidine preferentially from the other R/Y mispair. Thus, if all ACs were corrected to GC (R excised) and all GTs were corrected to GC (Y excised), there could be complete disparity in favor of GC($m$), with a $d$ of $\sim -0.25$ if no repair failure occurred.
For unlike R/R and Y/Y mispairs, disparity must arise by repair enzymes preferentially excising one purine from R/R and one pyrimidine from Y/Y mispairs. Strong disparity here has one cause: strong preferences in both types of mispair, both working in the same direction as regards + or m. For example, AG might correct largely to AT, giving + in the examples used here, with CT correcting largely to AT, also giving +. Low disparity has two causes: either low preferences in excision from the two mispairs or strong preferences working in opposite directions as regards + and m, e.g., AG correcting largely to AT (+) but CT correcting largely to CG (m).

Base Substitutions: Effects of Repair on Base Ratios

In the cases considered in the preceding section, whenever repair enzymes gave conversion direction disparity for +/m, there was an exactly corresponding amount of disparity for base pairs. Thus, if + has AT and m has GC at the point of heterozygosity, the G + C percentage at that site in the cross is initially 50%; total correction of AC and GT mispairs in h-DNA to GC(m) would give complete disparity in favor of m and GC, with 100% G + C at that site in products of conversion instead of the initial 50%. The exact correspondence of +/m (or A:a) disparity with base ratio disparity holds for all amounts of disparity, with or without some failure of correction (which equally reduces both types of disparity), because both types of disparity arise from the same correction event. Correction disparity for transitions and for transversions giving unlike R/R and Y/Y mispairs will change G + C percentage, but if disparity is possible for transversions giving like R/R and Y/Y mispairs (e.g., AA and TT from AT and TA), then it will change AT/TA or CG/GC ratios but not the G + C percentage.

In most natural populations, there will be many sites of heterozygosity, and by acting over many sites and many generations, gene conversion could slowly change base ratios from correction disparities. Although AT has been the base pair in + in the chosen examples, AT could equally well be a mutant base pair. Thus, depending on the base pair in a wild-type gene at a particular site, transitions from AT to GC could be + mutating to m at some sites but would be m mutating to + at other sites. If most AC mispairs in resulting h-DNA were corrected to GC and most GTs were corrected to GC, then one would get strong disparity for AT/GC, with more GCs than ATs, changing the G + C percentage in both cases but in the first case (AT = +) with disparity favoring m and in the second case (AT = m) with disparity favoring +. The disparity-producing properties of repair enzymes could therefore give systematic changes in G + C percentage but with disparity sometimes favoring + and sometimes favoring m, depending on the original frequencies of particular base pairs in wild-type genes. Systematic changes in base ratio could therefore occur whether the number of sites with disparity favoring wild-type sites was equal or unequal to the number of sites with disparity favoring mutant sites, and both conditions occur in data analyzed by Lamb and Helmi (1982) and Lamb (1984).

Asymmetric h-DNA Formation with Chromatid Preferences

Consider the case where the + chromatid in a + × m cross has an AT base pair, the m chromatid has a corresponding CG base pair, and strands 1 and 3 have the same polarity. If there is no strand preference in asymmetric h-DNA formation
Conversion Affects Base Ratios and Total DNA

(i.e., strands 1 and 3 invade a homologue as often as do 2 and 4), chromatid preference in invasion gives equal frequencies of the two types of mispair but changes the base ratio by h-DNA formation. For example, if all chromatid invasions were of m invaded by + but half of the invasions were by strand 1 of + (carrying A) and half were by strand 2 of + (carrying T), the initial base ratio would be changed from 1 AT (in +):1 CG (in m) to 1 AT (unchanged in +):½AG + ½CT (invasion products in what was m). The initial base ratio of 1A:1T:1C:1G has become 1½A:1½T:½C:½G, changing the G + C percentage. Such changes in base ratio from asymmetric h-DNA will also give exactly corresponding amounts of +/m (or A/a) disparity, with 1+:1m changing to 1½+:½m at the h-DNA stage in this example.

The actual disparities observed for base pairs and +/m in such a case would depend on whether there was disparity in the direction of correction as well as asymmetric h-DNA chromatid preference and on whether the two types of disparity worked in the same direction (e.g., both favoring +) or in opposing directions. Unlike disparity from correction-enzyme preference in repair, disparity from asymmetric h-DNA chromatid preference is unlikely to cause systematic changes in base ratios because the preference is unrelated to the base pair involved, so that gains of AT from CG at one point could be counterbalanced by gains of CG from AT at another point. The lack of systematic effects from asymmetric h-DNA with chromatid preference also applies to heterozygous frame-shifts, since the chromatid preference could equally favor the strand with the extra base or the strand without it. It is possible that there could be asymmetric h-DNA formation with chromatid preference and with strand preference. One could then get unequal numbers of the two sorts of mispairs expected, but once again one would get equal amounts of +/m and base ratio disparity, as in the case in which there is no strand preference.

Discussion

It is clear that long-term trends in base ratios and DNA amounts could be partly controlled by the specificity of enzymes concerned in the correction of mispairs or nonpairs in h-DNA. For example, changes in whether a correction enzyme (or enzyme system) preferentially cuts the looped-out or the nonlooped strand of a nonpair could decrease DNA content of a genome, thereby saving metabolic energy, or increase DNA content, thereby increasing evolutionary potential. The rate of such changes must depend on the amounts of polymorphism and heterozygosity and on conversion frequencies and the amounts of disparity in conversion direction. Changes in DNA content from frame-shifts will also depend on the relative frequencies of addition and deletion mutations and on the lengths of those changes, as considered above.

For base substitutions, +/m or A/a disparity from conversion is usually clear evidence of a matching disparity that changes base ratios as measured by G + C percentage, but for some transversions it is evidence of AT/TA or CG/GC base ratio changes. This is true whether disparity arises from correction or from asymmetric h-DNA formation with chromatid preference.

As shown in the data in tables 1 and 2, many organisms have a G + C percentage that departs considerably from 50%. If an organism had, for example, a high G + C percentage in coding sequences, then a GC or CG base pair would be wild type more often than was AT or TA. Preferential correction of mispairs AC,
AG, TC and TG, to GC or CG rather than to AT or TA, would mean that
corrections to wild type for base substitutions were more common (averaged over
many sites) than conversions to mutant. An AT-rich coding region could, however,
show preferential conversion to mutant. It would be informative to get data on
relative frequencies of disparity favoring wild type or mutant in organisms of
different G + C percentage, in genome regions within an organism that differ in G
+ C percentage, and for mutations with known base changes.

As shown in table 3, frame-shifts appear to convert preferentially to mutant,
with negative mean \( d \) values for all five loci, but this probably arises because most
of these frame-shifts are ICR-170-induced additions, with correction enzymes
preferentially cutting the nonlooped strand in nonpairs. There is a lesser tendency
for base substitutions to convert preferentially to wild type, with positive mean \( d \)
values for four out of five loci. Much more extensive data and knowledge of the
molecular nature of the mutations are required to assess whether frame-shifts have
different overall trends in correction direction compared with base substitutions. It
is clear, however, from data analyzed by Lamb (1984), that frame-shifts tend to
have more frequent correction of nonpairs than base substitutions have for mispairs,
and the higher disparities observed for frame-shifts might thus change DNA amounts
faster than base ratios would be changed by base-substitution disparity. Counterbal-
ancing that would be the expected higher average selection coefficients against
frame-shifts, which usually give complete loss of function in a coding sequence,
whereas base substitutions are sometimes cryptic (synonymous or tolerated; see
Lamb [1975]).

It is not known how many correction enzymes there are, whether they are
mispair or nonpair specific, or whether they are regionally specific. The transfection
experiments referred to in the Introduction suggest that the base preferentially
excised depended on the nature of the mismatch and was sometimes a purine,
sometimes a pyrimidine. The study by Dohet et al. (1984) on \( \lambda \) repair in transfected
Escherichia coli showed different efficiencies of repair of the eight possible mispairs,
with significant disparities in direction of correction for all mispairs, four favoring
a + allele and four favoring an \( m \) allele. Transfection experiments thus show that
correction disparity occurs in prokaryotes and can be extreme, even where the two
strands have similar amounts of methylation.

The reasoning that disparity indicates changes in base ratios and total DNA
amounts is unaffected by whether recombination in a chromosome is initiated by
single-strand breaks, as in the Holliday (1964) and Meselson and Radding (1975)
models, or by a double-strand break, as in the model of Szostak et al. (1983), since
the latter model also explains disparity by h-DNA formation and correction.

The interactions of gene conversion, mutation, selection, dominance, and
genetic drift were considered by Lamb and Helmi (1982). The effects of those
interactions on base ratios and total DNA amounts will be similar to their effects
on allele frequencies, because of matching amounts of disparity for \(+m\) and for
base ratios and for cutting looped/nonlooped strands, as discussed above. The
amount of selection will obviously depend on whether the sites are in coding or
noncoding regions of DNA, e.g., increases in DNA amounts resulting from disparity
in correction for frame-shifts are much more likely to survive in noncoding regions
than in coding regions. Changes in G + C percentage could also occur particularly
in noncoding regions, and through cryptic mutations, tolerated and favorable
substitutions in coding regions. Selection for G + C percentage could operate at several levels: (1) effects of base changes on the amino acid sequence and function of particular polypeptides, (2) effects of G + C content on the thermal stability of duplexes, and (3) the possibly greater UV sensitivity of low G + C-percentage DNA, because of its having more adjacent thymines, which more easily form dimers than other pyrimidine combinations. Large changes in G + C percentage in coding regions might change the organism's overall amino acid composition, with selection operating on such broader changes as well.

Disparity in gene conversion could provide a mutationally controllable mechanism by which base ratios and total DNA amounts could be gradually changed, sometimes advantageously, independently of mutation frequencies, which are themselves subject to selection and genetic control. A change in disparity from a mispair-repair enzyme should eventually lead to a new equilibrium value for base ratios, with a new balance between conversion, mutation, selection, and other factors.

Acknowledgement

I thank Mr. S. Zwolinski for critically reading the manuscript.

LITERATURE CITED


RICHARD B. FLAVELL, reviewing editor

Received September 5, 1984.
Information for Contributors

Molecular Biology and Evolution is a bimonthly journal devoted to the interdisciplinary science between molecular biology, and evolutionary biology. The journal emphasizes 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 normally written by solicited authors 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 (addresses below). Any manuscript or any part of a manuscript which 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 addressed 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 except under the written permission of the copyright owner. The journal will not be responsible for the loss of manuscripts at any time.

Publication is taken to imply that the authors are prepared to make available to the public any unpublished sequences on which the paper is based and any clone of cells, DNA, or antibodies used in the experiments reported. This principle also applies to computer programs.

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 and manuscript editor's notes and queries. 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.
Information for Contributors

All organisms mentioned must be identified by their scientific binomials 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 [1966]:61–76). 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.).

As recommended to the IUB, 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 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.)

For the style of citations, please note the following examples:

Journal articles:

Information for Contributors

Books:


Book chapters:


The abbreviations of periodicals should be those used by the Council of Biology Editors. Periodical titles may also be written out.

Articles should include the name of the reviewing editor (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 4¼ inches wide by 7⅞ 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

**Editor in Chief**
Walter M. Fitch
Department of Physiological Chemistry
University of Wisconsin—Madison
1300 University Avenue
Madison, Wisconsin 53706

**Managing Editor**
Masatoshi Nei
Center for Demographic and Population Genetics
University of Texas Health Science Center at Houston
P.O. Box 20334
Houston, Texas 77225

**Associate Editors**
Roy J. Britten
Kerckhoff Marine Laboratory
California Institute of Technology
Corona del Mar, California 92625

Wesley M. Brown
Division of Biological Sciences
University of Michigan
Ann Arbor, Michigan 48109

Richard B. Flavell
Department of Cytogenetics
Plant Breeding Institute
Cambridge CB2 2LQ
England

Ken W. Jones
Department of Genetics
University of Edinburgh
West Mains Road
Edinburgh EH9, 3JN Scotland

Richard K. Koehn
Department of Ecology and Evolution
State University of New York at Stony Brook
Stony Brook, New York 11794

Robert K. Selander
Department of Biology
University of Rochester
Rochester, New York 14627
If the key word is genetics—

behavioral genetics  molecular genetics  biochemical genetics
clinical genetics and  population genetics  cytogenetics
  counseling  cancer genetics  mutagenesis/somatic
  immunogenetics  developmental genetics  cell genetics

—the key research is in

The American Journal of Human Genetics

Established in 1948, the AJHG is one of the leading journals addressing central issues in contemporary genetics. The Journal—which is the official publication of The American Society of Human Genetics—provides a record of research and review relating to heredity in humans and to the application of genetic principles in medicine, psychology, anthropology, and the social sciences as well as in related areas of molecular and cell biology.

Editor: Dr. David E. Comings.

Unlock the information you need:

Published bimonthly by The University of Chicago Press. 1-year rates: Individuals $75; Institutions $95. Outside of the U.S.A. postage surcharges vary according to destination. Visa and MasterCard accepted. Mail complete charge card information, payment, or purchase order to The University of Chicago Press, Journals Division, P.O. Box 37005, Chicago, IL 60637. Please note: Subscriptions entered on a calendar-year basis only. Students accepted as members of the society only. For student and all other membership rates, write to the Executive Office, American Society of Human Genetics, P.O. Box 6015, Rockville, MD 20850.

Subscribe to the AJHG.
That’s how many pages we publish each year—original research, review articles, serial features, clinical briefs, medical perspectives, symposia and special articles—all devoted to increasing your understanding of infectious disease.

The Journal of Infectious Diseases is the foremost outlet for new research in infectious diseases—diseases that are responsible for more than 20% of hospital admissions and 50% of first office or clinic visits. Journal articles communicate information from clinical, epidemiologic, and laboratory experience. They provide a basis for understanding the causes, pathogenesis, diagnosis, and treatment of infections as well as the biology of the host response. Published monthly. Martha Yow, editor.

Reviews of Infectious Diseases, JID's companion journal, fulfills a twofold purpose: first, to publish review articles offering comprehensive consideration of recent literature in a particular area or on a group of related topics; and second, to publish thematic studies focusing on central topics in infectious diseases. Often most or all of an issue is devoted to papers treating a single theme. Published bimonthly. Frequent supplements. Edward H. Kass, editor.

Sponsored by the Infectious Diseases Society of America.

One-year subscriptions: JID/RID combination: Institutions $225; Individuals $105. Special rates for students, interns, residents, and fellows as well as some society members. For further information, write to Circulation Manager, University of Chicago Press, P.O. Box 37005, Chicago, IL 60637 USA.
An authoritative international journal published monthly for the European Molecular Biology Organisation

- Manuscripts typeset and published within 12 weeks of acceptance
- Large American quarto format with papers printed on high quality art paper
- No page charges
- Issues mailed by air to all non-European subscribers
- Reduced subscription rates for individuals
- Listed in Current Contents/Life Sciences

The EMBO Journal is published monthly for the European Molecular Biology Organization and provides rapid publication for reports of significant original research in molecular biology and related areas. More specifically, papers cover biochemistry, cell and developmental biology, molecular genetics of eukaryotes, prokaryotes and organelles, methodology, immunology, structural biology, neurobiology and virology. Only research reports of the highest quality are published and access to the journal is essential for molecular biologists everywhere.

A free sample copy, containing full 'instructions to authors', is available on request.

Editor in Chief B. Hirt, Chairman of EMBO Council
Executive Editor John Toole, Executive Secretary of EMBO

Editorial Board

Subscription rates for Volume 3 (12 issues in 1984): libraries £155.00/US$310.00; individuals whose library subscribes £50.00/US$100.00; EMBO members £40.00/US$80.00; USA price valid in Canada and the Americas. Pound Sterling (£) definitive elsewhere

IRL Press, PO Box 1, Eynsham, Oxford OX8 1JY, UK
IRL Press, Suite 907, 1911 Jefferson Davis Highway, Arlington, VA 22202, USA
this publication is available in microform

Please send me additional information.

Name ________________________________
Institution ____________________________
Street ________________________________
City ________________________________
State ___________ Zip ________________

University Microfilms International
300 North Zeeb Road
Dept. P.R.
Ann Arbor, MI 48106
U.S.A.

18 Bedford Row
Dept. P.R.
London, WC1R 4EJ
England
Where the animals roam . . .

Physiological Zoology

From chloride transport in freshwater mussels to digestive efficiencies of wild howler monkeys, from sweat gland secretion in the black Bedouin goat to respiratory physiology of freely diving harbor seals — PZ remains one of the few journals concerned with whole organisms and animal populations as well as with cellular mechanisms.

Offering increasing numbers of comparative physiologists and physiological ecologists the research that is shaping the field, PZ's scope encompasses all aspects of physiology:

- water and ions
- respiration
- temperature
- metabolism
- biochemistry

Physiological Zoology, now bimonthly. Edited by C. Ladd Prosser and James E. Heath.

One-year rates: Individuals $38.00; Institutions $65.00; Students with copy of ID $25.00. Outside USA add $3.00. Visa or MasterCard accepted. Send complete charge card information, purchase order, or payment to The University of Chicago Press, P.O. Box 37005, Chicago, Illinois 60637.

THE UNIVERSITY OF CHICAGO PRESS
Forthcoming

Nearly Identical Allelic Distributions of Xanthine Dehydrogenase in Two Populations of *Drosophila pseudoobscura*
Tim P. Keith, Lisa D. Brooks, Richard C. Lewontin, Juan Carlos Martinez, and Donna Larson Rigby

The Adaptation of Enzymes to Temperature: Catalytic Characterization of Glucosephosphate Isomerase Homologues Isolated from *Mytilus edulis* and *Isognomon alatus*, Bivalve Molluscs Inhabiting Different Thermal Environments
John G. Hall

Evolutionary Change of Restriction Cleavage Sites and Phylogenetic Inference for Man and Apes
Masatoshi Nei and Fumio Tajima

Gene Regulatory Divergence among Species Estimated by Altered Developmental Patterns in Interspecific Hybrids
Henry R. Parker, David P. Philipp, and Gregory S. Whitt
Molecular Biology and Evolution

Editor in Chief
Walter M. Fitch, University of Wisconsin

Managing Editor
Masatoshi Nei, University of Texas

Associate Editors
Roy J. Britten, California Institute of Technology
Wesley M. Brown, University of Michigan
Richard B. Flavell, Cambridge University
Ken W. Jones, University of Edinburgh
Richard K. Koehn, State University of New York at Stony Brook
Robert K. Selander, University of Rochester

Editorial Assistant
Jon Bartels

Editorial Board
Fred W. Allendorf
University of Montana
Norman Arnheim
SUNY at Stony Brook
John C. Avise
University of Georgia
Francisco J. Ayala
University of California, Davis
Jaap Beintema
Rijksuniversiteit Groningen
Anthony Hugh D. Brown
CSIRO, Canberra
L. L. Cavalli-Sforza
Stanford University
Elizabeth A. Craig
University of Wisconsin—Madison
Irving P. Crawford
University of Iowa
W. Ford Doolittle
Dalhousie University
Gabriel Dover
Cambridge University
Marshall H. Edgell
University of North Carolina
George Fox
University of Houston
Morris Goodman
Wayne State University
R. Grantham
Université de Lyon
Michael Grunstein
University of California, Los Angeles
Barry G. Hall
University of Connecticut
Daniel L. Hartl
Washington University
Alec Jeffreys
University of Leicester
Motoo Kimura
National Institute of Genetics, Japan

Costas Krimbas
Agricultural College of Athens, Greece
Charles H. Langley
National Institute of Environmental Health Sciences, North Carolina
Richard C. Lewontin
Harvard University
Wen-Hsiung Li
University of Texas at Houston
Alan Maxam
Harvard Medical School
George L. Gabor Miklos
Australian National University
Roger Milman
University of Iowa
Takashi Miyata
Kyushu University, Japan
Tomoko Ohta
National Institute of Genetics, Japan
Dennis A. Powers
Johns Hopkins University
Nils Ryman
University of Stockholm, Sweden
Barbara Schaal
Washington University
Temple F. Smith
Northern Michigan University
Howard M. Temin
University of Wisconsin—Madison
Michael S. Waterman
University of Southern California
Sherman Weissman
Yale University Medical School
Gregory S. Whitt
University of Illinois at Urbana-Champaign
Eleutherios Zouros
Dalhousie University
Molecular Biology and Evolution
Volume 2, Number 2, March 1985

87 Estimating Divergence Times of Drosophila Species from DNA Sequence Comparisons (Letter to the Editor)
Simon Easteal and John G. Oakeshott

92 Biochemical Pathways in Prokaryotes Can Be Traced Backward through Evolutionary Time (Review)
Roy A. Jensen

109 A Comparative Description of Mitochondrial DNA Differentiation in Selected Avian and Other Vertebrate Genera
Louis G. Kessler and John C. Avise

127 Tempo and Mode of Concerted Evolution in the L1 Repeat Family of Mice
Sandra L. Martin, Charles F. Voliva, Stephen C. Hardies, Marshall H. Edgell, and Clyde A. Hutchison III

141 Coevolution of the Glucose Dehydrogenase Gene and the Ejaculatory Duct in the Genus Drosophila
Douglas R. Cavener

150 A New Method for Estimating Synonymous and Nonsynonymous Rates of Nucleotide Substitution Considering the Relative Likelihood of Nucleotide and Codon Changes
Wen-Hsiung Li, Chung-I Wu, and Chi-Cheng Luo

175 The Effects of Mispair and Nonpair Correction in Hybrid DNA on Base Ratios (G + C Content) and Total Amounts of DNA
B. C. Lamb
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, 5801 Ellis Avenue, Chicago, Illinois 60637, and sponsored by the Molecular Biology and Evolution Society and the American Society of Naturalists.

Subscription Rates U.S.A.: institutions, 1 year $60.00; individuals, 1 year $30.00. Student subscription rate, U.S.A.: 1 year $24.00 (copy of student ID must accompany subscription). Other countries add $4.00 for each year's subscription to cover postage. Single copy rates: institutions $10.00, individuals $5.00. 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. Editorial correspondence should be addressed to the Editor in Chief of Molecular Biology and Evolution, 528A Service Memorial Institute, University of Wisconsin, 1300 University Avenue, Madison, Wisconsin 53706. 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., Operation Center, P.O. Box 765, Schenectady, New York 12301. 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 the publisher.

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

© 1985 by The University of Chicago. All rights reserved.