A Comparison of the Small Ribosomal RNA Genes from the Mitochondrial DNA of the Great Apes and Humans: Sequence, Structure, Evolution, and Phylogenetic Implications

James E. Hixson and Wesley M. Brown

Division of Biological Sciences and Museum of Zoology, University of Michigan

Restriction endonuclease fragments produced by EcoRI/AvaI or KpnI digestion and containing the small (12S) ribosomal RNA (rRNA) genes from the mitochondrial DNAs (mtDNAs) of the common chimpanzee, pygmy chimpanzee, gorilla, and orangutan were inserted into the plasmids pBR322 or pADD1. After species verification the inserted fragments were digested with SauIIIA, subcloned into M13mp7 vectors, and sequenced. The small rRNA gene sequences were compared with each other and with the published human sequence (Anderson et al. 1981). Substitutions were detected at 118 of the 955 nucleotide positions compared. Pairwise, the sequence differences ranged from 1% (between the chimpanzee species) to 9% (comparisons involving the orangutan); the proportion that were transitions ranged from 87% to 100%. Deletions and/or additions were noted at seven locations. With respect to evolutionary sequence lability, kinetic analysis indicated the presence of at least two classes of nucleotide positions; the more labile class occurs in sequences thought to form self-complementary duplexes (stems) in the mature rRNA. The high frequency of compensating substitutions, which maintain base-pairing within these sequences, corroborates their inferred structure. Phylogenetic inferences drawn from the sequence comparisons support the notion of an approximately equidistant relationship among chimpanzees, gorilla, and man, with the orangutan much less closely related. However, inference from a shared deletion suggests that the gorilla and the chimpanzees may be more closely related to one another than they are to man.

Introduction

Mammalian mitochondrial DNA (mtDNA) contains 37 genes that encode either proteins or functional RNA products (Anderson et al. 1981, 1982; Bibb et al. 1981; reviewed by Attardi 1985; Brown 1985). Among the latter are the small (12S) and large (16S) ribosomal RNA (rRNA) genes whose products form the structural RNA components of the small and large mitochondrial ribosomal subunits, respectively (Vesco and Penman 1969; Eperon et al. 1980). The rRNA gene sequences in vertebrate mtDNAs do not appear to be as highly conserved in evolution as their nuclear counterparts. Evidence from both DNA heteroduplex thermostability analysis and restriction-endonuclease cleavage-map comparisons has indicated that these mitochondrial

1. Key words: substitution rate, transition bias, deletions, primates, phylogeny.

Address for correspondence and reprints: Dr. Wesley M. Brown, Division of Biological Sciences, University of Michigan, Ann Arbor, Michigan 48109.

2. Current address: Southwest Foundation for Biomedical Research, Department of Genetics, P.O. Box 28147, San Antonio, Texas 78284.

© 1986 by The University of Chicago. All rights reserved.
0737-4038/86/0301-3104$02.00
rRNA genes change rapidly, although less rapidly than the mitochondrial genes that code for proteins (Dawid 1972; Brown et al. 1979; Brown and Simpson 1981; Ferris et al. 1981a, 1981b; Cann et al. 1984). However, precise quantitative estimates for the rates of change of mitochondrial rRNA genes have not been obtained, nor are there data from which the kinetics of rRNA gene evolution may be derived.

Secondary-structure models for the large and small rRNAs of prokaryotes, eukaryotes, chloroplasts, and mitochondria have been proposed and subsequently refined; these models are based on both comparative sequencing studies and on structural analyses (see Brimacombe et al. 1983; Maly and Brimacombe 1983; Woese et al. 1983 for reviews). Despite the vast evolutionary distances separating the origins of prokaryotes, eukaryotes, and eukaryotic cellular organelles, the major features of the proposed secondary structures of their rRNAs are strikingly similar, and much of their primary sequence is also conserved (Woese et al. 1980, 1983; Kuntzel and Kochel 1981; Noller and Woese 1981; Stiegler et al. 1981a, 1981b; Zwieb et al. 1981; Anderson et al. 1982; Maly and Brimacombe 1983; Gray et al. 1984; Spencer et al. 1984). However, the differences in the structures and sequences of the rRNAs from these groups of organisms and organelles are equally striking. These differences are reflected in the dramatically different sizes of the rRNAs between prokaryotes, eukaryotes, chloroplasts, and mitochondria and also between mitochondrial rRNAs from different major taxonomic groups (e.g., single-celled eukaryotes, plants, and animals). The mitochondrial rRNAs of animals are especially small. Although many descriptive studies have been performed and more are in progress, most of the interesting and important questions about the functional and evolutionary consequences of the rRNA structural similarities and differences remain to be addressed.

In order to add generally to the fund of knowledge about rRNA genes and specifically to investigation of some issues and questions noted above, we have cloned and sequenced both the small and the large rRNA genes from the mtDNAs of the four species of great apes: common and pygmy chimpanzee, gorilla, and orangutan. Together with man and the gibbons, these form a closely related group, the hominoida, that for both molecular and evolutionary reasons is extremely well suited for comparative studies (discussed on p. 226 of Brown et al. 1982). The most strongly supported phylogeny for these species and their closest relatives is shown in figure 1. The mitochondrial rRNA gene sequences from the great apes, when compared to each other and to the sequences of human mitochondrial rRNA genes reported by Anderson et al. (1981), provide a kinetic picture of ribosomal gene evolution over a divergence-time range that has not been previously explored. In this paper we report the sequences and the results of the sequence comparisons for the small rRNA genes. Those for the large rRNA genes are nearly complete and will be reported separately (J. E. Hixson, L. L. Szura, and W. M. Brown, work in progress).

Material and Methods
DNA Preparation

Mitochondrial DNAs were prepared from frozen tissues of the lowland gorilla (Gorilla gorilla), pygmy chimpanzee (Pan paniscus), and orangutan (Pongo pygmaeus). Frozen tissues were provided by Dr. O. Ryder of the San Diego Zoo. Common chimpanzee (Pan troglodytes) mtDNA was prepared from the cultured cells originally described by Benirschke et al. (1974), using the same culture conditions employed by Brown and Vinograd (1974) for the growth of cells from the woolly monkey (Lagothrix
FIG. 1.—Evolutionary relationships among hominoids (apes and man) and their nearest relatives, the old-world monkeys. The precise relationship among chimpanzees, gorilla, and man is presently disputed and is therefore shown as an unresolved trichotomy. Estimates of hominoid divergence times are also disputed; these are based on values suggested by Zihlman et al. (1978), Gingerich (1984, 1985), Hasegawa et al. (1984), and Sibley and Ahlquist (1984). Abbreviations: CC = common chimpanzee; PC = pygmy chimpanzee; GO = gorilla; HO = man; OR = orangutan; GB = gibbon; OWM = old-world monkeys; MYBP = millions of years before the present.

cana). Plasmid DNAs were prepared according to a protocol based on the method of Clewell and Helinsky (1969). The viral and replicative (RF) forms of M13 bacteriophage DNA were isolated as described by Sanger et al. (1980).

Mitochondrial DNA Cloning

Purified mtDNA was digested with either KpnI (gorilla and orangutan) or with both EcoRI and AvaI (common and pygmy chimpanzee). These digests were mixed, respectively, with KpnI-digested, alkaline phosphatase-treated pADD1 plasmid DNA (Greenberg et al. 1983) or with EcoRI/AvaI-digested, phosphatase-treated pBR322 plasmid DNA (Bolivar et al. 1977), after which bacteriophage T4 DNA ligase was added. The ligated mixtures were used to transform Ca&-treated Escherichia coli (strain HB101) cells (Mandel and Higa 1970) by a standard protocol (Maniatis et al. 1982). Plasmid DNAs were prepared from ampicillin-resistant, tetracycline-sensitive colonies by the procedure of Birnboim and Doly (1979) and screened for restriction fragments of the correct size, as determined for these species by earlier studies (Ferris et al. 1981a, 1981b).

Subcloning into M13 DNA and Sequencing

After amplification and preparation of the recombinant plasmid DNAs, the mitochondrial portions were separated from the pADD1 or pBR322 portions by digestion with KpnI or EcoRI/AvaI, respectively, followed by preparative agarose gel electrophoresis. The mitochondrial fragments were electroeluted from the gels, digested with SauIIIA, mixed with BamHI-digested M13mp7 RF DNA (Messing et al. 1981), and incubated with T4 DNA ligase. After transformation of E. coli (strain JM103) with the ligated DNA, colonies exhibiting recombinant (white) phenotypes were purified, cultured, and used as sources of single-stranded viral DNAs. These DNAs were screened for the presence of mtDNA sequences. The insert sizes were characterized and the inserts sequenced using the methods of Sanger et al. (1977, 1980). The complete sequences of inserted mtDNA fragments that were longer than 400 bp were obtained
by constructing stepwise deletions using the DNase I deletion protocol described by Hong (1982).

Enzymes and Other Reagents

All enzymes were purchased from commercial suppliers. Restriction endonucleases and T4 DNA ligase were from New England Biolabs (NEB), bacterial alkaline phosphatase and DNase I from Worthington Biochemicals, and calf intestinal alkaline phosphatase from Boehringer-Mannheim (B-M). The large (Klenow) fragment of DNA polymerase I was from New England Nuclear, B-M, and NEB. Alpha-\(^{32}\)P-deoxynucleotide triphosphates (3,000 Ci/mmol; 10 mCi/ml) were from Amersham (Chicago). Dideoxynucleotides and deoxynucleotides were from P-L Biochemicals. All chemicals used were reagent grade, except for cesium chloride, which was technical grade.

Phylogenetic Analysis of Sequence Data

The aligned sequences were compared by a method that computes the minimum number of nucleotide substitutions required by each of all possible branching orders among the taxa. Only those nucleotide positions that differ among one or more of the taxa provide information, since positions at which all taxa are identical cannot differentiate among them. By parsimony, the branching order(s) requiring the least number of substitutions is(are) considered most likely to reflect the actual evolutionary relationships among the taxa. Positions at which only one species differs have been included because such positions provide information about the amount of change that its sequence has undergone since it last shared a common ancestor with other members of the group—and thus they also provide information about rates of change. The computations were performed by means of an AMDAHL 5860 computer using version 2.1 of PAUP (Phylogenetic Analysis Using Parsimony), which was written and adapted for the computer by Dr. J. Swofford (Illinois Natural History Survey, Champaign, Ill.).

Results and Discussion

Cloning and Sequencing of the Small rRNA Genes

Following the preparation of the mtDNAs from frozen tissues or cultured cells, the identity of each was confirmed by comparative restriction-endonuclease analysis using enzymes that produce species-diagnostic mtDNA fragment patterns (Ferris et al. 1981a, 1981b). In both common and pygmy chimpanzee mtDNAs the small rRNA gene occupies an internal position in a 5,000-bp EcoRI/AvaI fragment (fig. 2). The products of EcoRI/AvaI digestion of mtDNA from these two species were cloned into the unique site created by EcoRI/AvaI digestion of the plasmid pBR322 (Bolivar et al. 1977). In gorilla and orangutan mtDNAs the small rRNA gene occupies an internal position in 2,843- and 3,092-bp KpnI fragments, respectively. The products of KpnI digestion of mtDNA from each of these species were cloned into the unique KpnI site of the plasmid pADD1 (Greenberg et al. 1983). After transformation of Escherichia coli HB101 and selection on antibiotic medium, resistant colonies were screened for plasmids containing inserts of \(\sim 5,000\) bp (EcoRI/AvaI) or 3,000 bp (KpnI). The species identities of the inserts were reconfirmed by comparative restriction analysis.

Recombinant plasmids were digested with SauIIIA, and fragments containing mtDNA sequences were subcloned into the BamHI site of the bacteriophage vector M13mp7 (Messing et al. 1981), from which they were sequenced using the method
FIG. 2.—Cleavage maps, cloning and sequencing strategies, and gene content in the region containing the small (12S) ribosomal RNA gene of common and pygmy chimpanzee (CC and PC, respectively), gorilla (GO), and orangutan (OR) mtDNA. As in other mammalian mtDNAs, the small rRNA gene (SrRNA) is flanked on one side by the tRNA^Phe gene (F) and the control region (which contains the D-loop) and on the other side by the genes for tRNA^Val (V) and the large rRNA (LrRNA) (J. E. Hixson, L. L. Szura, and W. M. Brown, unpublished data). KpnI or EcoRI/AvaI fragments were used for primary cloning, and Sau3A fragments for subcloning into M13mp7 sequencing vectors. Lengths and directions of sequencing runs from Sau3A sites are indicated by plain arrows, and those from internal sites (i.e., from deletion points) by arrows with triangular butts. Most arrows represent several sequencing runs. The numbers indicate the distance, in base pairs, from the start of the small rRNA gene, and increase in the direction of rRNA transcription. Distances to the KpnI, EcoRI, AvaI, and most distal Sau3A sites are approximate. The short vertical arrows represent Sau3A sites. The immediately adjacent genes for cytochrome b (cyt b), tRNA^Tyr (T), tRNA^Pro (P), and unidentified reading frame 1 (URF 1), some of which contained sites that were useful in cloning, are also shown.

of Sanger et al. (1977) (fig. 2). As indicated in figure 2, several of the longer fragments were sequenced in a stepwise manner using the DNase deletion technique of Hong (1982). The L-strand sequences (which correspond to those of the rRNAs) are presented in figure 3, along with the human small rRNA gene sequence reported by Anderson et al. (1981).

Sequence Alignment

In figure 3 the complete small rRNA gene sequence of common chimpanzee mtDNA is presented in alignment with those of the pygmy chimpanzee, gorilla, man, and orangutan. Alignments were uncomplicated because of the low level of sequence divergence among the small rRNA genes of these closely related species. Although a few small (1–4-bp) deletions and additions have occurred, these do not lead to ambiguities in alignment except between the gorilla and common chimpanzee sequences on the 310–320 interval of figure 3, where a 4-bp deletion in a polycytidine (poly-C) tract has occurred in each sequence. These deletions probably arose independently, because corresponding deletions do not occur in the pygmy chimpanzee or man, the two species most closely related to the common chimpanzee and gorilla (fig. 1). Thus
Fig. 3.—Aligned sequences of the mitochondrial small rRNA genes from the great apes and man. The entire common chimpanzee (CC) sequence is shown. For the other taxa, only nucleotides that differ from the CC sequence are shown. A hyphen (-) indicates that a space has been introduced to maintain alignment. Sequences postulated to form base-paired structures (stems) in the rRNA are enclosed by lines. Although
### Table 1

**Pairwise Comparisons of the 12S rRNA Gene Sequences among the Great Apes and Man**

<table>
<thead>
<tr>
<th>COMPARISON</th>
<th>SUBSTITUTIONS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SP 1/SP 2 bp</td>
<td>% TS</td>
</tr>
<tr>
<td>CC/PC</td>
<td>946</td>
<td>1.2</td>
</tr>
<tr>
<td>CC/GO</td>
<td>944</td>
<td>4.1</td>
</tr>
<tr>
<td>CC/HO</td>
<td>948</td>
<td>3.7</td>
</tr>
<tr>
<td>CC/OR</td>
<td>945</td>
<td>9.2</td>
</tr>
<tr>
<td>PC/GO</td>
<td>946</td>
<td>4.0</td>
</tr>
<tr>
<td>PC/HO</td>
<td>950</td>
<td>3.2</td>
</tr>
<tr>
<td>PC/OR</td>
<td>947</td>
<td>8.9</td>
</tr>
<tr>
<td>GO/HO</td>
<td>949</td>
<td>3.5</td>
</tr>
<tr>
<td>GO/OR</td>
<td>946</td>
<td>9.0</td>
</tr>
<tr>
<td>HO/OR</td>
<td>951</td>
<td>8.8</td>
</tr>
</tbody>
</table>

**NOTE.**—The human 12S rRNA gene sequence is from Anderson et al. (1981). In calculating the number of base pairs, positions at which one or both species lacked a nucleotide were not counted. Abbreviations: SP = species; CC = common chimpanzee; PC = pygmy chimpanzee; GO = gorilla, HO = man; OR = orangutan; bp = no. of base pairs in the comparison; TS = transitions; TV = transversions.

* Estimated number of pairwise differences from addition and/or deletion events.

The positions shown in figure 3 for these 4-bp deletions have been designated rather than determined. In the gorilla and man, the alignment of the T's at position 314 is also ambiguous. The exact positions to which both the T's and the 4-bp deletions should be assigned cannot be determined from the DNA sequences or from considerations of small rRNA secondary structure.

**Sequence Divergence of the Small rRNA Genes**

The great ape and human small rRNA genes have diverged at 118 of their 955 positions via nucleotide substitutions, at 5 locations via small (1–4-bp) deletions, at one (nos. 225–227) via either an addition or a deletion of 3 bp, and at one (shown as no. 884) via a single nucleotide addition (fig. 3). Although these differences are spaced throughout the small rRNA gene, the 5' half is somewhat less conserved and contains 1.6 times as many differences as are found in the 3' half. The kinds and amounts of substitutions and the number of deletions and additions observed in pairwise comparisons among the species are presented in table 1. In agreement with the evolutionary relationships postulated for these species (fig. 1), the small rRNA gene sequences are least divergent (1.2%) between the pygmy and common chimpanzee; more divergent (3.2%–4.1%) between the chimpanzee, gorilla, and man; and most divergent (8.8%–9.2%) in pairwise comparisons involving the orangutan.

The most notable feature of the substitutions is the predominance of transitions, which comprise 87%–94% of the nucleotide differences in pairwise comparisons among A:C, R:R, and Y:Y pairs were excluded from the analysis of the CC sequence, we note that models and empirical evidence for hydrogen-bonded A:A and A:C pairs exist (see de Bruijn and Klug 1983). The baselines of the enclosures are numbered in accordance with the secondary-structure model shown in fig. 5. Sequences with the same (primed and unprimed) numbered baselines constitute the two strands of a given stem. Arrows at the beginning or end of a given line of sequences indicate that a stem region continues on the line of sequences immediately above or below it, respectively. Species abbreviations are those given in figs. 1 and 2. The human sequence is from Anderson et al. (1981).
the chimpanzee, gorilla, and orangutan and 100% of those between the common and pygmy chimpanzees (table 1). The predominance of transitions in mitochondrial sequence comparisons between closely related taxa has been observed for protein and tRNA genes among the hominoidea (Brown et al. 1982), for the cytochrome oxidase subunit II gene between rat species (Brown and Simpson 1982), for portions of the URF1 and cytochrome oxidase subunit I genes between zebra species (Higuchi et al. 1984), and for the D-loop region among seven humans (Aquadro and Greenberg 1983; Greenberg et al. 1983). Among the hominoidea (Brown et al. 1982) the magnitude of the transition bias was inversely related to divergence time. The present results do not suggest such a relationship. It is likely, however, that this arises in part from constraints imposed by secondary structural considerations, such as helix (i.e., stem) conservation, and in part to stochastic fluctuation (the number of transversions in this sample is small; see table 1). This interpretation is supported by the observation that the extreme bias toward transitions observed in these comparisons does not persist in mitochondrial rRNA gene sequence comparisons between more distantly related taxa (e.g., mouse, cow, and man; Anderson et al. 1981, 1982; Bibb et al. 1981; Brown 1985).

One addition and six deletions also contribute to the divergence of the small rRNA genes. In figure 3 the addition of a C in the pygmy chimpanzee sequence is shown at position 884. However, the actual position is uncertain, and the best resolution achievable is within the 884–888 interval, which consists of a tract of five C's. The deletions at position 67 in both chimpanzee sequences, at positions 60–61 in the pygmy chimpanzee, and between positions 310 and 320 in the gorilla also occur in poly-C tracts. The common chimpanzee deletion shown from 316–319 differs slightly from these in that it includes an A that is present at position 319 in the other sequences as well as three C's. Although homopolymer repeats of other nucleotides are found in the small rRNA gene (poly-A tracts occur at a greater frequency than poly-C tracts), these were not associated with addition or deletion events.

Poly-C tracts associated with small deletions have been observed in the noncoding region of human mtDNA (Aquadro and Greenberg 1983; Greenberg et al. 1983) and also occur adjacent to extensive deletions in the noncoding region of gorilla mtDNA (J. E. Hixson, D. R. Foran, and W. M. Brown, manuscript in preparation). The exact role of these tracts in the generation of deletions in mtDNA is as yet unclear, although as repeated sequences they could serve as deletion sites via slipped mispairing during mtDNA replication (Streisinger et al. 1966; Albertini et al. 1982). The association of poly-C tracts with small deletions is curious and may be of significance. As noted by Zwieb et al. (1981), poly-C tracts occur at locations in the mitochondrial small rRNA gene that correspond to sites with greatly elaborated secondary structures in both prokaryotic and eukaryotic cytoplasmic rRNAs (see insets A' and B' of fig. 5 below).

Evolutionary Dynamics of Small rRNA Genes

We observed substitutions at 12.4% of the 955 nucleotide positions in the small rRNA genes (fig. 3). This is one-half the percentage that was observed in an 896-bp mitochondrial tRNA and protein gene sequence from these same species (Brown et al. 1982). In pairwise comparisons, the percentages of substitutions among chimpanzees, gorilla, and man are threefold lower and those with the orangutan twofold lower than in the protein and tRNA gene comparison (Brown et al. 1982). In contrast, when the small rRNA genes from more distantly related species are compared (e.g., mouse,
cow, and man), the divergence values (22%–25%) are only slightly lower than those found in the protein and tRNA gene sequence study (30%–33%) and are nearly identical to the divergence values found for the most highly conserved mitochondrial protein genes (Bibb et al. 1981; Anderson et al. 1982; Brown 1985).

We noted the disparity between the amounts of divergence observed in rRNA genes when compared among closely versus more distantly related species in an earlier study (Ferris et al. 1981a) and speculated that the ribosomal genes might accumulate substitutions at a rate that is linear with time but intermediate to the rates of silent (fast) and replacement (slow) substitutions in mitochondrial protein genes. When a greater range of divergence times is examined, however, the kinetics of the small rRNA genes are clearly nonlinear, and must be bi- or multiphasic (curve B, fig. 4). Qualitatively, the kinetics resemble those inferred from cleavage-map comparisons of mammalian mtDNAs (Brown et al. 1979) and from sequence comparisons of an 896-bp portion of hominoid mtDNAs (curve A, fig. 4; also see Brown et al. 1982). The differences in the amounts of substitution observed between the rRNA and protein genes of closely related species are accounted for by an approximately twofold difference in their initial rates of substitution (compare curves A and B, fig. 4). The kinetics of substitution in the small rRNA genes, like those in the mitochondrial protein genes, is due to the existence of two (or more) classes of sites. These data clearly indicate that no class of sites as unconstrained as the silent sites of protein genes is present. Instead, the initial (most rapid) rate of substitution takes place at positions where change is constrained by a requirement for the maintenance of secondary structure (e.g., in stem regions) but not by the primary sequence per se. The slower rate occurs mostly at positions in regions whose primary sequences must be under direct selection. The sequences of these regions are highly conserved between mammalian orders (curve

Fig. 4.—Percentage difference vs. divergence times for pairwise sequence comparisons of genes and portions of genes from mammalian mtDNAs. Curve A, portions of two protein genes; curve B, small rRNA genes; curve C, a nonconsecutive 301-bp sequence of the small rRNA gene that is highly conserved (>80% identity) relative to the small rRNA gene sequence of *Escherichia coli* (Zwieb et al. 1981). Divergence-time estimates are those used in fig. 1. The dashed portions of curves B and C are interpolations. Curve A is based on data from Brown et al. (1982) and, for the point at 25 Myr before the present, on unpublished data for the vervet monkey (*Cercopithecus aethiops*) (L. L. Szura and W. M. Brown). The reference *E. coli* small rRNA sequences used to derive curve C are from Brosius et al. (1978).
Hominoid Mitochondrial rDNA Gene Sequences

1 I

C, fig. 4; also see Anderson et al. 1982) and among all prokaryotic and eukaryotic small-subunit rRNAs (Stiegler et al. 1981a; Zwieb et al. 1981; Woese et al. 1983; Gray et al. 1984; Spencer et al. 1984).

Structural Considerations

The small rRNA RNAs of animal mitochondria can be folded into a structure that retains many of the same intrastrand interactions that are characteristic of small-subunit rRNAs from prokaryotic, chloroplast, and eukaryotic cytoplasmic ribosomes (Kuntzel and Kochel 1981; Stiegler et al. 1981a; Zwieb et al. 1981; Woese et al. 1983). However, the small rRNA of mammalian mitochondria is smaller than that of E. coli by ~500 bp and accordingly lacks some of the structural features postulated for E. coli rRNA, as shown in insets A' and B' of figure 5. The probable correspondences among several of the structural features of primate mitochondrial and bacterial small-subunit rRNAs are listed in table 2. In figure 5 we present a folding pattern for the small rRNA RNAs of the common chimpanzee. This pattern does not differ significantly from that presented for the small rRNA RNAs of human mitochondria by Zwieb et al. (1981) as modified by Maly and Brimacombe (1983). The precise locations of the sequence changes among the ape and human species relative to this pattern are given in figure 3.

Long-range interactions within the small rRNA RNAs of ape and human mitochondria (stems 2, 3, 10, and 17 in fig. 5) allow it to be divided into four domains similar to those that characterize the small rRNA gene of E. coli (Stiegler et al. 1981a, 1981b; Maly and Brimacombe 1983; Woese et al. 1983). Each domain consists of a series of stem-and-loop structures that are interconnected by single-stranded regions. The long-range interactions that determine the domains are highly conserved. The stem structures within domains appear to be more variable in their extent of conservation. Stem structures are formed by intrastrand base pairing between different regions of the rRNA. When nucleotide substitutions occur in a region that forms one strand of a stem, the stability of that stem can be maintained only if compensating substitutions occur in the region that forms the other strand of the stem. The occurrence of compensating substitutions at a frequency greater than expected from chance is evidence that intrastrand base pairing has been maintained by selection and implies that the stems are functionally important structures. As indicated by the data in table 2, most
Table 2
Differences in Stem Regions of Mitochondrial (mt) Small-Subunit rRNA Genes Relative to
the Common Chimpanzee rRNA Gene Sequence

<table>
<thead>
<tr>
<th>HELIX NO. (Escherichia coli/mt)*</th>
<th>COMPENSATED b</th>
<th>UNCOMPENSATED b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PC</td>
<td>GO</td>
</tr>
<tr>
<td>3/3</td>
<td>...</td>
<td>2</td>
</tr>
<tr>
<td>4/4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>7/6</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>11/7</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>15/8</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>18/9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>20/11</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>23/13</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>24/14</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>25/15</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>32/21</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>42/26</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>44/28</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>45/29</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

NOTE.—Abbreviations are those given in table 1.

a mt small-subunit rRNA helices are numbered as in figs. 3 and 5. The corresponding E. coli helices are numbered according to fig. 2 of Maly and Brimacombe (1983).
b Numbers presented are differences relative to the common chimpanzee sequence only and do not usually correspond to the absolute number of differences within a given helix.
c The E. coli helix corresponding to mt helix no. 24 could not be determined with certainty.

of the interspecific differences observed in regions postulated to form stems are com-

pensated.

Conversely, the presence of several uncompensated substitutions (e.g., in stems 6 and 7 [Domain I], 24 [Domain III], and 28 [Domain IV] of fig. 5) might suggest that the stems that contain them are not functionally relevant. However, it is worth considering the possibility that stems with lower thermodynamic stability may be required at some ribosomal locations to accommodate conformational changes that must occur during protein synthesis.

Phylogenetic Analysis of the Sequences

Figure 1 depicts the hypothesis for the phylogenetic relationships among the hominoidea that appears best supported by data from both molecular and nonmo-

lecular studies (see, e.g., Goodman 1963; Sarich and Wilson 1967; Wilson and Sarich 1969; Ferris et al. 1981a, 1981b; Andrews and Cronin 1982; Brown et al. 1982; Wolpoff 1982; Templeton 1983a; Pilbeam 1984; Sibley and Ahlquist 1984; Benveniste 1985; Martin 1985; Ueda et al. 1985). The relative branching order of the chimpanzee, gorilla, and human lineages, shown as an unresolved trichotomy in figure 1, is currently disputed among evolutionary biologists. Based on an extensive statistical analysis of our data (Ferris et al. 1981b; Brown et al. 1982), Templeton (1983a, 1983b) has argued strongly that the chimpanzee and gorilla lineages are most recently diverged. Based on their own extensive DNA:DNA hybridization studies, Sibley and Ahlquist (1984)
have argued with equal conviction that the chimpanzee and human lineages are most recently diverged. Whatever the relative branching order among these lineages, the high degree of resolution obtainable with both the mtDNA sequence data (Brown et al. 1982; this study; J. E. Hixson, L. L. Szura, and W. M. Brown, unpublished results) and the DNA:DNA hybridization data (Sibley and Ahlquist 1984) indicate that the separation in time between these respective divergence events was probably small. In all but minor details the phylogenetic hypothesis shown in figure 1 appears to be highly supported by many studies, using diverse kinds of data and methods of data analysis. However, Kluge (1983) and Schwartz (1984) have recently championed alternative branching orders that differ radically from this (figs. 6A and 6B, respectively).

Although the small rRNA gene sequence data cannot conclusively resolve these disputes, they can be used to evaluate some of these alternative hypotheses. The minimum number of substitution steps was computed for each of all possible branching orders for the five taxa (see Material and Methods). The most parsimonious branching orders (figs. 6C and 6D) each required 133 independent substitution events. The next most parsimonious (fig. 6E) required 134 events. The third most parsimonious, that shown in figure 1, required 138 events, indicating that the chimpanzee-gorilla-human trichotomy, although corroborated by our present knowledge, is unlikely to be true in a literal sense. Without an outgroup sequence (one from a gibbon or siamang would be most appropriate), the networks shown in figures 6C–6E cannot be empirically rooted. Thus, formally, the branching order in figure 6E could be topologically equivalent either to that in 6A or to that in 6B. However, the branching orders depicted in figures 6A and 6B required 4.5 times as many events along the lineage leading from the common divergence node in figures 6C–6E to the orangutan than they do along the lineage leading from this node to man. Although differences in branch length characterize all branching orders in figure 6, the ratio between the extremes (maximum/minimum) is much smaller (1.24–2.06) in the branching orders shown in figures 6C–6E. Among equally parsimonious outcomes, we regard those with extreme rate

![Fig. 6.](image)

**FIG. 6.**—Two alternative hypotheses of hominoid relationships (A–B) and the three most parsimonious branching relationships (C–E) based on sequence comparisons of the mitochondrial small rRNA genes. Branching orders A and B correspond, respectively, to hypotheses advanced most recently by Kluge (1983) and Schwartz (1984). The common node in C–E that joins the orangutan lineage to those leading to the remaining taxa (see text) is indicated by the heavy dot. The small and large numbers for C–E indicate, respectively, the individual branch lengths and the total (minimum) number of substitutions required to generate each.
differences to be less likely than their more uniform counterparts, unless there is independent evidence in support of the extreme rate fluctuations. No such evidence exists among the very extensive molecular data available for these taxa, and the present data thus support a version of the branching order shown in figure 1 in which the human-chimpanzee-gorilla trichotomy is resolved (i.e., one of the branching orders shown in figs. 6C–6E).

Another way that data from sequence comparisons may be analyzed for phylogenetic inference is based on the assumption that the evolutionary distance between species is proportional to the magnitude of their sequence difference. The evolutionary distances between the hominoid species are thus proportional to the percentage of divergence of their rRNA genes (table 1). The tree that best fits these distances has the same branching order (but different branch lengths) as that shown in figure 6C. This is also the best fitting tree when the data from our previous mtDNA sequence comparisons (Brown et al. 1982) are analyzed in this way (Nei et al. 1985; Nei and Tajima 1985). However, the difference in the fit of trees in figures 6C–6E is not statistically significant, and the alternative branching orders represented by the trees in figures 6D and 6E cannot be excluded (Nei et al. 1985).

Given the conflicting "best" trees produced by these different methods, the lack of statistical significance between "best" and "next best" trees, and the lack of agreement about the validity and evolutionary significance of the methods themselves (Farris 1981; Templeton 1983a, 1983b; Nei and Tajima 1985), it is not reasonable to draw firmer conclusions on the basis of comparisons of the mitochondrial small rRNA gene sequences.

Phylogenetic Analysis of Additions and Deletions

The seven additions and deletions (located at positions 60–61, 67, 73, 225–227, 310–313, 316–319, and 884 in fig. 3) can also be used to infer phylogeny. Five of these additions/deletions are restricted to single species and do not provide information about phylogeny. The other two are single-base-pair deletions, neither of which occurs in a homopolymer repeat. Because of the small number of events in the rRNA gene sequences, we assume that the probability of the same base pair being deleted independently in different lineages is low. Thus, the deletion shared by both chimpanzee species at position 67 was most probably acquired from a common ancestor, and the deletion event probably occurred in this ancestor after its divergence from the lineages leading to the other hominoid species. A similar argument may be made for the deletion shared among the two chimpanzee species and the gorilla at position 73. If this interpretation is correct, the deletion supports the branching order shown in figure 6E. This is the same branching order that was "best" in the parsimony analysis of tRNA and protein gene sequences among hominoid mtDNAs (Brown et al. 1982). However, because the data presently consist of rRNA gene sequences from only one individual of each species, the possibilities (1) that this deletion might be polymorphic within one or more of the species or (2) that it could simply represent convergence (with position 73 being a deletion "hotspot") must also be regarded as reasonable. A third possibility—that the deleted and nondeleted sequences may stem from chance segregation of a polymorphism in the species ancestral to man, the gorilla, and the chimpanzees and that it thus may not be indicative of phylogeny—has also been suggested (Nei 1985 and personal communication).
Acknowledgments

We thank Dr. Oliver Ryder and the Research Department of the San Diego Zoo for providing us with primate tissues from their frozen collection; L. Lynne Szura and John Kluge, Craig Moritz, and Ellen Prager for carefully reviewing the manuscript; and Margaret van Bolt and David Bay for assistance in preparing and photographing the figures. This work was supported in part by U.S. Public Health Service (USPHS) grants GM30144 (to W.M.B.) and RR07050 (Biomedical Research Support Grant to the Vice President of Research of The University of Michigan) and by Horace H. Rackham and Phoenix Memorial Project grants (to W.M.B.) from The University of Michigan. J.E.H. received support from USPHS training grant GM07544 and from a University of Michigan Cancer Research Institute Fellowship.

LITERATURE CITED


BROWN, G. G., and M. V. SIMPSON. 1982. Novel features of animal mtDNA evolution as shown


Masatoshi Nei, reviewing editor

Received June 6, 1985; revision received August 20, 1985.
Conservation of a Long Open Reading Frame in Two *Neurospora* Mitochondrial Plasmids

*Frank E. Nargang*

Department of Genetics, University of Alberta

The nucleotide sequence of a specific region of the mitochondrial plasmid from the *Neurospora intermedia* Varkud-lc strain was determined. Analysis of the sequence revealed the presence of a long (up to 710 amino acids) ORF. This ORF is almost identical to a previously characterized ORF in the mitochondrial plasmid from the *Neurospora crassa* Mauriceville-lc strain. When the ORFs from the two plasmids are compared over their entire length of 2,133 bp, only 34 nucleotide substitutions are found (greater than 98% identity). These substitutions result in only nine amino acid replacements in the protein sequences predicted from the two ORFs. Though no function can be assigned to the putative products of these ORFs, their high conservation of nucleotide and deduced amino acid sequence suggest that they are under selective pressure, presumably to preserve the function of some protein.

**Introduction**

Mitochondrial plasmids are small circular or linear DNAs that have been found in the mitochondria of various fungi (Collins et al. 1981; Stohl et al. 1982; Tudzynski et al. 1983; Garber et al. 1984; Natvig et al. 1984) and higher plants (Pring et al. 1977, 1982; Boutry and Briquet 1982; Brennicke and Blanz 1982; Palmer et al. 1983). The mitochondrial plasmids of *Neurospora* have been found in several strains isolated from nature (Collins et al. 1981; Stohl et al. 1982; Natvig et al. 1984), though they have never been detected in common laboratory strains. Many of the *Neurospora* plasmids are quite similar and have been classified, on the basis of DNA hybridization experiments, into three major groups (Natvig et al. 1984).

The entire DNA sequence of the mitochondrial plasmid from the Mauriceville-lc strain of *Neurospora crassa* has been determined (Nargang et al. 1984). The sequence revealed many interesting structural features and led to the speculation that the Mauriceville plasmid may be related to mitochondrial introns. Perhaps the most remarkable feature of the Mauriceville DNA sequence is the presence of a long ORF that could potentially encode a hydrophilic protein consisting of as many as 710 amino acids. However, no protein product exclusive to the mitochondria of the Mauriceville strain has been detected to date. Two large ORFs have also been identified in the DNA sequence of the S-2 mitochondrial plasmid of maize (Levings and Sederoff 1983). Again, there are no known protein products that correspond to either of the S-2 ORFs, although minor polypeptides specific to the mitochondria of strains with the S cytoplasm have been detected (Forde and Leaver 1980). Thus, the role of mitochondrial plasmid ORFs remains enigmatic. The alpha-sen DNA of *Podospora* mi-
Mitochondria contains a reading frame that has been shown to correspond to an intron of the cytochrome c oxidase subunit-1 gene (Osiewacz and Esser 1984). However, the alpha-sen DNA is not a true mitochondrial plasmid, since it is derived from the standard mitochondrial genome of the organism.

The discovery of a mitochondrial plasmid in the *Neurospora intermedia* Varkud-lc strain, which, on the basis of DNA hybridization and restriction mapping data, is very similar to the Mauriceville plasmid (R. A. Akins, L. L. Stohl, D. Grant, F. E. Nargang, and A. M. Lambowitz, unpublished data), offered a means of testing whether or not the same long ORF was present in two different mitochondrial plasmids. The presence of the same ORF in these two plasmids, which are found in two different *Neurospora* species (Perkins et al. 1976) isolated from geographically remote locations (Mauriceville, Texas and Varkud, India) would be a very strong argument for the conservation of the ORF sequence—and, by inference, of its function. A similar argument for conservation of function was applied to the unassigned reading frames (URFs) found in three mammalian mitochondrial genomes (Anderson et al. 1981, 1982; Bibb et al. 1981), and it has now been shown that these URFs do in fact give rise to mitochondrial proteins (Chomyn et al. 1983; Oliver et al. 1983; Michael et al. 1984).

The data described in the present report show that the ORF sequence in the Varkud mitochondrial plasmid is almost completely conserved relative to that in the Mauriceville mitochondrial plasmid.

**Material and Methods**

**Strains**

The Varkud strain of *Neurospora intermedia* (Fungal Genetics Stock Center no. 1823) was obtained from A. Lambowitz (St. Louis University). The isolation of the Varkud and Mauriceville strains carrying the mitochondrial plasmids discussed herein—and their classification into separate species based chiefly on crossing behavior—have been described by Perkins et al. (1976). Growth of *Neurospora* on solid and in liquid medium was as described elsewhere (Davis and de Serres 1970). The *Escherichia coli* strain HB101 (Bolivar and Beckman 1979) was used for propagation of cloned mitochondrial plasmid DNA in bacterial vectors. Strain JM103 (Messing et al. 1981) served as the host for transfections with M13 phage derivatives.

**Isolation of Mitochondria and Mitochondrial Plasmid DNAs**

Mitochondria were isolated from liquid culture by the flotation-gradient method (Lambowitz 1979). Mitochondrial plasmid DNA was isolated basically as described elsewhere (Collins et al. 1981), with minor changes. In brief, purified mitochondria from 3–5 liters of late-log-phase liquid culture were suspended in 5 ml of 50 mM Tris-HCl, pH 7.5, containing 5 mM ethylenediaminetetraacetic acid and then lysed by the addition of 0.5 ml 10% sodium dodecyl sulfate. CsCl (0.98 gm/1 ml lysate) and ethidium bromide (50 µl/1 ml lysate) were added, and the solution was spun for 8–12 h in a VTi65 rotor (Beckman) at 54,000 rpm. The lower band, containing mostly mitochondrial plasmid DNA, was collected and rebanded. Ethidium bromide was then removed by means of four extractions with isopropanol followed by dialysis to remove CsCl. This procedure yielded relatively pure mitochondrial plasmid DNA contaminated by small amounts of mtDNA.
Recombinant Plasmids and Phages, Transformation, and Isolation of Plasmid DNA

Plasmid pV2 was obtained from L. Stohl, D. Grant, and A. Lambowitz (St. Louis University). The plasmid was constructed by cloning BglII-digested Varkud mitochondrial plasmid into the BamHI site of the bacterial plasmid pBR322. The pV2 plasmid was used as the source of DNA for cloning into phage M13 derivatives for all of the sequence studies, with the exception of the M13 clone whose sequence traverses the BglII site (see fig. 2A below). The latter clone was constructed using purified Varkud mitochondrial plasmid DNA. Phage clones for dideoxy DNA sequence determination were constructed using the replicative form of the M13 phage derivatives mp18 and mp19 (Norrander et al. 1983). The procedures for transformation and isolation of plasmid DNA were as described elsewhere (Nargang et al. 1983).

Determination of DNA Sequence and Its Analysis

Determination of the DNA sequence using the dideoxy technique, computer analysis, and storage of the data were as described elsewhere (Nargang et al. 1983).

Results

The nucleotide and amino acid sequence of the region of the Varkud mitochondrial plasmid that was sequenced and that corresponds to the Mauriceville ORF region is shown in figure 1. The DNA sequence was translated using the genetic code for Neurospora mitochondria (UGA = trp, Heckman et al. 1980). Figure 2A shows a partial restriction map of the region sequenced and the strategy used. Both strands of the DNA were sequenced entirely. The first nucleotide in the single BglII site of the plasmid near the amino terminus of the long ORF is arbitrarily designated as position 1 of the DNA sequence. The numbering system downstream of the BglII site is thus consistent with that published for the ORF region of Mauriceville (Nargang et al. 1984). However, it should be noted that minor differences in the sequence and size of the two plasmids (Akins et al., unpublished data) cause the numbering system of nucleotides outside the ORF to vary between the two plasmids. For this reason and because the complete sequence of the Varkud plasmid has not been determined, the negative numbering system upstream of the BglII site has been employed. Translation of the nucleotide sequence shown in figure 1 revealed an ORF of as many as 710 amino acids if the first ATG, at position -90 to -88, is considered as the starting point of translation. A subset of smaller ORFs derived from the 710-amino acid ORF also exists, depending on which of the 14 internal ATG codons is considered as a possible translation starting site.

Figure 2B gives the position of nucleotide substitutions observed in the Varkud ORF when it is compared to the previously described Mauriceville ORF. A total of 34 differences are observed in the 2,178 positions compared. The ORF regions of the two plasmids are 98.6% identical. The substitutions appear to be randomly distributed over the codons of the ORF. Six differences are apparently the result of transversions, and 28 (82%) the result of transitions. Five substitutions have occurred in the first position of a codon, four in the second position, and 25 in the third position. Also shown in figure 2B is the position within the ORF of each nucleotide difference that results in an amino acid replacement with respect to the Mauriceville ORF. Of the 34 total differences, 24 (71%) are silent, whereas 10 alter the amino acid specified. One codon beginning at nucleotide 259 on figure 1 (change 8; table 1) is affected by two nucleotide changes, so that only nine amino acids differ between the long ORFs
FIG. 1.—The DNA sequence of the coding strand of the Varkud mitochondrial plasmid’s long ORF region. The amino acids encoded are also shown. The numbering system is as described in the text.
of Mauriceville and Varkud (table 1). Of the nine amino acid differences observed in the two ORFs, five still belong to the same chemical group of amino acids; thus, only four amino acids are radically different—namely, those changes designated as 23, 28, 30, and 32 in table 1—which are specified by the codons beginning at nucleotides 1,249, 1,609, 1,741, and 2,020, respectively, in figure 1. Interestingly, the latter three of these relatively severe changes—lys to glu, glu to val, and gly to asp—are found in the last 20% of the ORF sequence, near the carboxy terminus.

Discussion

Previous sequence data on the mitochondrial plasmid from the *Neurospora crassa* Mauriceville strain revealed the existence of a long ORF that covered more than 50% of the length of the plasmid (Nargang et al. 1984). Although no protein product peculiar to any mitochondrial plasmid has been identified to date, the existence of such a long ORF argues that it probably encodes a function that is under selective pressure. This
FIG. 2.—A, Partial restriction-endonuclease map and DNA sequencing strategy for the ORF region of the Varkud mitochondrial plasmid. Arrows show the direction and extent of the sequence determined from the indicated restriction site. Arrows starting at sites not indicated on the figure begin at AluI restriction sites. The first possible ATG codon and the TAA stop codon of the ORF are also indicated. B, Linear representation of the ORF (horizontal bar) relative to the region sequenced. Vertical lines within the horizontal bar indicate the sites of nucleotide differences between this and the Mauriceville mitochondrial plasmid ORF (Nargang et al. 1984). Vertical lines extending above the horizontal bar show those sites that result in amino acid replacements in the Varkud ORF relative to the Mauriceville ORF.

argument is enhanced by the observation that the putative protein product could indeed be translated from the major transcript of the plasmid (Nargang et al. 1984). The mitochondrial plasmid from the *Neurospora intermedia* Varkud strain is now the second *Neurospora* mitochondrial plasmid shown to contain a long ORF. As expected from the DNA hybridization and restriction-endonuclease mapping data the Varkud ORF is virtually identical to that of Mauriceville and like the Mauriceville ORF it could also be expressed from the plasmid's major transcript (Akins et al., unpublished data). The extent of the reading frame is entirely conserved in the two plasmids, and the few nucleotide differences that do exist are confined to nucleotide substitutions rather than being deletions or additions that would alter the frame. This contrasts with the situation outside the ORF region, where numerous and more substantial differences between the two plasmids are found (Akins et al., unpublished data). Since the ORF makes up more than 50% of the sequence in each plasmid, it is unlikely that it is so well conserved by chance alone. The distribution and type of amino acid changes observed between the two ORFs also suggest conservation from a functional viewpoint; that is, most of the nucleotide differences are silent or result in relatively innocuous amino acid changes (table 1). Those changes that cause rather drastic amino acid differences are clustered near the carboxy terminus of the ORF and may reflect the relative unimportance of that region to the function of the putative ORF protein. Although all these observations argue strongly for conservation of the function of the protein, no hint as to its actual role exists as yet. The most likely possibilities are that either (1) the protein is required for the maintenance of the plasmid itself or (2) there exists a function that provides some advantage to the host strain in a particular environment. A region with eight out of 12 identical amino acids between the Mauriceville ORF sequence and that of the putative polymerase of cauliflower mosaic virus was noted previously (Nargang et al. 1984), and a distant relationship of the ORF to retroviral reverse transcriptases has recently been detected (Michel and Lang 1985).

It is not known whether the 98.6% identity observed between the ORF regions of the Varkud and Mauriceville plasmids is different than that to be expected between
Table 1
Summary of Nucleotide Differences as They Affect Codons in the Long Open Reading Frame of the Mauriceville and Varkud Mitochondrial Plasmids

<table>
<thead>
<tr>
<th>MITOCHONDRIAL PLASMID</th>
<th>CODON DIFFERENCE NO.</th>
<th>NO. OF FIRST NUCLEOTIDE IN CODON</th>
<th>Mauriceville Codon</th>
<th>Amino Acid</th>
<th>Varkud Codon</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1*</td>
<td>25</td>
<td>ATA</td>
<td>ile</td>
<td>ATG</td>
<td>met</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>AAC</td>
<td>asn</td>
<td>AAT</td>
<td>asn</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>55</td>
<td>GGA</td>
<td>gly</td>
<td>GGG</td>
<td>gly</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>61</td>
<td>GTA</td>
<td>val</td>
<td>GTC</td>
<td>val</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>187</td>
<td>AAA</td>
<td>lys</td>
<td>AAG</td>
<td>lys</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>220</td>
<td>ATT</td>
<td>ile</td>
<td>ATC</td>
<td>ile</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>238</td>
<td>GAG</td>
<td>glu</td>
<td>GAA</td>
<td>glu</td>
<td></td>
</tr>
<tr>
<td>8*</td>
<td>259</td>
<td>ATC</td>
<td>ile</td>
<td>CTG</td>
<td>leu</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>280</td>
<td>TAC</td>
<td>tyr</td>
<td>TAT</td>
<td>tyr</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>415</td>
<td>CAC</td>
<td>his</td>
<td>CAT</td>
<td>his</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>427</td>
<td>CCT</td>
<td>pro</td>
<td>CCC</td>
<td>pro</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>460</td>
<td>CTA</td>
<td>leu</td>
<td>TTA</td>
<td>leu</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>499</td>
<td>GAT</td>
<td>asp</td>
<td>GAC</td>
<td>asp</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>505</td>
<td>AAA</td>
<td>lys</td>
<td>AAG</td>
<td>lys</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>622</td>
<td>ACC</td>
<td>thr</td>
<td>ACT</td>
<td>thr</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>766</td>
<td>ACA</td>
<td>thr</td>
<td>ACG</td>
<td>thr</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>844</td>
<td>TTC</td>
<td>phe</td>
<td>TTT</td>
<td>phe</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>913</td>
<td>CCC</td>
<td>pro</td>
<td>CCT</td>
<td>pro</td>
<td></td>
</tr>
<tr>
<td>19*</td>
<td>925</td>
<td>GTC</td>
<td>val</td>
<td>ATC</td>
<td>ile</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1,012</td>
<td>GGA</td>
<td>gly</td>
<td>GGG</td>
<td>gly</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>1,033</td>
<td>AAT</td>
<td>asp</td>
<td>AAC</td>
<td>asp</td>
<td></td>
</tr>
<tr>
<td>22*</td>
<td>1,168</td>
<td>GTT</td>
<td>val</td>
<td>GCT</td>
<td>ala</td>
<td></td>
</tr>
<tr>
<td>23*</td>
<td>1,249</td>
<td>ATA</td>
<td>ile</td>
<td>ACA</td>
<td>thr</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>1,372</td>
<td>ACT</td>
<td>thr</td>
<td>ACC</td>
<td>thr</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>1,483</td>
<td>TAT</td>
<td>tyr</td>
<td>TAC</td>
<td>tyr</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>1,564</td>
<td>AAG</td>
<td>lys</td>
<td>AAA</td>
<td>lys</td>
<td></td>
</tr>
<tr>
<td>27*</td>
<td>1,606</td>
<td>ATG</td>
<td>met</td>
<td>TTG</td>
<td>leu</td>
<td></td>
</tr>
<tr>
<td>28*</td>
<td>1,609</td>
<td>AAG</td>
<td>lys</td>
<td>GAG</td>
<td>glu</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>1,729</td>
<td>GGC</td>
<td>gly</td>
<td>GGA</td>
<td>gly</td>
<td></td>
</tr>
<tr>
<td>30*</td>
<td>1,741</td>
<td>GAG</td>
<td>glu</td>
<td>GTG</td>
<td>val</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>1,987</td>
<td>CGC</td>
<td>arg</td>
<td>CGT</td>
<td>arg</td>
<td></td>
</tr>
<tr>
<td>32*</td>
<td>2,020</td>
<td>GGT</td>
<td>gly</td>
<td>GAT</td>
<td>asp</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>2,023</td>
<td>AAA</td>
<td>lys</td>
<td>AAG</td>
<td>lys</td>
<td></td>
</tr>
</tbody>
</table>

NOTE.—Asterisks indicate nucleotide substitutions causing amino acid replacements. The nucleotide numbering system refers to fig. 1.

normal mitochondrial genes in different species of *Neurospora*, since no data on mtDNA sequence divergence between species are available for the organism. Although at least six size differences between the standard mtDNAs of the Varkud and Mauriceville strains have been detected by restriction-endonuclease analysis (Collins and Lambowitz 1983), the relationship between this size variation and the variation observed in the mitochondrial plasmids of the two strains is unclear, since significant structural diversity of mtDNAs is commonly observed both within and between species of *Neurospora* (Bernard et al. 1976; Manella et al. 1979; Collins and Lambowitz 1983) as well as in other lower eukaryotes (Sederoff 1984).
Considerable data are available on sequence divergence and evolution of animal mtDNAs. As discussed by Brown (1983), these studies have shown that the more closely related the species, the higher the overall ratio of transitions to transversions and the higher the ratio of silent to replacement substitutions in protein-coding genes. The nucleotide differences between the Mauriceville and Varkud plasmids in the ORF region are mostly transitions (28 of 34 = 82%) and mostly silent with respect to amino acid substitutions (24 of 34 = 71%). It has been suggested that the preponderance of transitions observed in the mtDNAs of closely related animal systems reflects a bias in the mutational process (Brown et al. 1982; Brown and Simpson 1982; Aquadro and Greenberg 1983). It is tempting to speculate that a similar bias exists in Neurospora mitochondria; however, in the absence of more data, particularly sequence comparisons between noncoding regions, no conclusions can be reached.

Given the obvious differences between animals and Neurospora, the data on animal mitochondrial systems cannot be realistically employed to extrapolate a reliable evolutionary distance between the Mauriceville and Varkud mitochondrial plasmids. In fact, it has already been noted that mtDNAs of higher-order plants are evolving at a slower rate than those of animals (Bonen et al. 1984). However, in light of the observations described above, it is probably reasonable to assume that the Mauriceville and Varkud mitochondrial plasmids are fairly closely related. The fact that the codon beginning at nucleotide 1,259 in figure 1 (difference number 8; table 1) contains changes in both the first and third positions is of interest with regard to the evolutionary relationship of the two plasmids; that is, all the other amino acid differences between the two ORFs can be explained as being the result of one-step events, but difference number 8 requires at least two distinct steps.

Natvig et al. (1984) have discussed the possibilities that the Neurospora mitochondrial plasmids have either been retained in various species as the species diverged or been introduced since those species diverged. Consideration of the geographical distance between the locations where the Varkud and Mauriceville strains were isolated makes the separate introduction of different forms of this mitochondrial plasmid into these two (which would make the sequences xenologous; Gray and Fitch [1983]) seem less likely than the codivergence of the species and the plasmids (which would make the sequences orthologous). Unfortunately, it is not known whether either these or related plasmids exist only in these Varkud and Mauriceville strains that happened to be picked from the wild or whether they have a wider distribution. It is therefore possible that many more forms of these plasmids have been fixed in various isolated populations.

Acknowledgments

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Alberta Cancer Board. I am grateful to Joy Correia and Daman Vig for technical assistance and to Kenneth Morgan for helpful discussion and for reviewing the manuscript.

LITERATURE CITED


BONEN, L., P. H. BOER, and M. W. GRAY. 1984. The wheat cytochrome oxidase subunit II gene has an intron insert and three radical amino acid changes relative to maize. EMBO J. 3:2531–2536.


NARGANG, F. E., J. B. BELL, L. L. STOHL, and A. M. LAMBOWITZ. 1983. A family of repetitive palindromic sequences found in Neurospora mitochondrial DNA is also found in a mitochondrial plasmid DNA. J. Biol. Chem. 258:4257–4260.


WESLEY M. BROWN, reviewing editor

Received April 26, 1985; revision received August 12, 1985.
The restriction enzyme TagI digests 0.2% of the genomic DNA from the grasshopper Caledia captiva to a family of sequences 168 bp in length (length of consensus sequence). The sequence variation of this "Taq family" of repeat units was examined among four races from C. captiva to assay the pattern of evolution within this highly repeated DNA. The Taq-family repeats are located in C-banded heterochromatin on at least one member of each homologous pair of chromosomes; the locations range from centromeric to telomeric. Thirty-nine cloned repeats isolated from two population IA individuals along with 11 clones from seven populations taken from three of the races demonstrated sequence variation at 72 positions. Pairwise comparisons of the cloned repeats, both within an individual and between different races, indicate that levels of intraspecific divergence, as measured by reproductive incompatibility, do not correlate with sequence divergence among the 168-bp repeats. A number of subsequences within the repeat remain unchanged among all 50 clones; the longest of these is 18 bp. That the same 18-bp subsequence is present in all clones examined is a finding that departs significantly ($P < 0.01$) from what would be expected to occur at random. Two other cloned repeats, from a reproductively isolated race of C. captiva, have sequences that show 56% identity with this 18-bp conserved region. An analysis showed that the frequency of occurrence of an Rsal recognition site within the 168-bp repeat in the entire Taq family agreed with that found in the cloned sequences. These data, along with a partial sequence for the entire Taq family obtained by sequencing uncloned repeats, suggest that the consensus sequence from the cloned copies is representative of this highly repeated family and is not a biased sample resulting from the cloning procedure. The 18-bp conserved sequence is part of a 42-bp sequence that possesses dyad symmetry typical of protein-binding sites. We speculate that this may be significant in the evolution of the Taq family of sequences.

Introduction

Numerous families of highly repeated DNA have been examined in plant and animal species (for reviews, see Appels and Peacock 1978; John and Miklos 1979). These studies have identified two major classes of highly repeated sequences, namely, those sequences that are interspersed within a genome and those that are tandemly repeated. The latter sequences are often referred to as satellite sequences because of the characteristic satellite band that they can form when centrifuged in certain buoyant
density gradients. These sequences are tandemly repeated as many as $10^6$ times within a genome and can be located on either a few or all of the chromosome pairs in a complement. The sequences are typically associated with heterochromatic regions of chromosomes. In addition, the sequence families that have been analyzed are often defined by the presence of a specific endonuclease recognition site located within each of the tandem repeats. The regular spacing of such sites within the tandemly arranged repeats suggests that homogeneity exists among the sequences within a given family. However, recent studies have revealed that extensive heterogeneity can be present within an apparently homogeneous family of repeats (Lee and Singer 1982; Lam and Carroll 1983a). This heterogeneity can be organized such that similar variant repeats are located on the same chromosome (Appels and Dvorak 1982; Lee and Singer 1982).

In the present study we have examined highly repeated sequences from the grasshopper *Caledia captiva*. This species was chosen because of the presence of extreme variability in the amount and cytological distribution of heterochromatin among the *C. captiva* races (Shaw et al. 1980). In addition, preliminary molecular analyses revealed sequence variation (this paper; M. L. Arnold, unpublished data) within a particular highly repeated sequence family.

**Material and Methods**

*Caledia captiva* Samples

The races and populations examined in the present study have been denoted by numerals and letters, respectively. Initially, on the basis of chromosomal variation, *C. captiva* was divided into four races (Moran and Shaw 1977) that were designated Southeast Australian (S.E.A.), Moreton, Torresian, and Daintree. Laboratory hybridization studies have demonstrated that the S.E.A. and Moreton races vary in their level of reproductive incompatibility, depending on which populations are utilized in the hybridization experiments (Shaw et al. 1980; Coates and Shaw 1984). Thus, F2 inviability in these crosses ranges from 0% to 42%. In addition, S.E.A. and Moreton individuals produce from 54% to 100% inviable F2 progeny when they are crossed with Torresian individuals (Shaw et al. 1980). Furthermore, crosses involving the Daintree race and the other three races result in either no progeny or completely sterile F1 individuals (Shaw et al. 1980). In the present paper the S.E.A., Moreton, Torresian, and Daintree races are designated 1, 2, 3, and 4, respectively. The following populations were used in the in situ and/or sequencing experiments: Gundaroo, Araluen, and Lakes Entrance (1A, 1B, and 1C, respectively); Scrubby Creek and Peregian (2A and 2B, respectively); Bongmuller, Insulator Creek, and Papuan *C. captiva* (3A, 3B, and 3C, respectively); and Daintree (4A).

**DNA Isolation**

Samples of DNA were isolated from individual, eviscerated grasshoppers and from pooled samples containing 10–40 individuals using the techniques of Appels and Dvorak (1982). Yields of DNA from individual grasshoppers ranged from ~80 μg to ~1,000 μg, depending on size.

**Cloning of the Highly Repetitive Sequence from *Caledia***

DNA renaturing at a Cot $<0.02$ (mol $\times$ sec)/liter was isolated from population 1A and used to synthesize radioactive probes ($^{32}$P c-DNA or $^3$H c-RNA) by standard procedures. A TaqI digest of total DNA from the 1A sample was probed with this DNA, and a fragment consisting of ~150 bp was shown to be a subset of the Cot
<0.02 fraction (fig. 1). This fragment was isolated from agarose gels by electroelution. These TaqI fragments were then inserted into the Clal site of pBR322 and used to transform E. coli ECR291; two of six plasmids showed significant hybridization to the Cot 0.02 fraction. Colonies were selected for inability to grow in the presence of 15 μg/ml of tetracycline while retaining their resistance to 200 μg/ml of ampicillin. A sample of clones sensitive to tetracycline was selected for individual plasmid preparations. Plasmid DNA from these clones was digested with TaqI, electrophoresed in a 2% agarose gel, and then transferred to Gene Screen (New England Nuclear). The filter was exposed to a 32P-nick-translated Cot 0.02 probe for 16–20 h in 50% formamide, 0.5% sodium dodecyl sulfate (SDS), 3 × SSC (0.45 m NaCl, 0.045 m Na3C6H5O7), and 5 × Denhardt's solution (0.1% bovine serum albumin, 0.1% ficoll, 0.1% polyvinylpyrrolidone). The filter was washed for at least 4 h in 2 × SSC/0.1% SDS, dried, covered with preexposed X-ray film, and placed at −70 C (Laskey 1980) to screen for hybridizing clones.

In Situ Hybridization

Mitotic cells were prepared on air-dried slides from 8–10-day-old embryos (Webb 1976) from population 1B. The chromosomal location of sequences homologous to the 150-bp cloned sequence was demonstrated by the hybridization of 3H-cRNA to these preparations following the procedure of Appels et al. (1978).

Saturation Hybridization

Saturation hybridization experiments to determine copy number of the repetitive sequence in population 1A followed the procedure of Bishop et al. (1969).

M13 Cloning and Dideoxy Sequencing of Taq-Family Repeats

A 100-μg sample of the plasmid pSEA1 was digested with the restriction enzyme TaqI. This sample was electrophoresed in a 2% agarose gel, and the band corresponding to the C. captiva DNA was cut out of the gel and subsequently electroeluted at 40 mA for 3–4 h. This DNA was then ligated into the Accl site of M13 mp8 (Messing 1981).

Taq-family repeats were isolated from two individuals from population 1A and from pooled grasshoppers for populations 1C, 2A, 2B, 3A, 3B, 3C, and 4A. The pro-

![Fig. 1. — Autoradiograph from Southern hybridization of TaqI-digested total population 1A DNA to a 32P c-DNA of the 168-bp repeat from population 1A.](image)
The procedure used to isolate the Taq-family repeats from these samples was identical to that used to isolate the repeat that provided the pSEA1 clone (see Cloning of the Highly Repetitive Sequence from Caledia). These sequences were cloned directly into the AccI site of M13 mp8, M13 mp9, or M13 mp10 (Messing 1981; Messing and Vieira 1982), without first cloning into the ClaI site of pBR322. This facilitated the screening of larger numbers of cloned sequences in order to isolate the Taq-family representatives from races 3 and 4 that contained 20% and 2%, respectively, of the number of copies in population 1 (M. L. Arnold and D. D. Shaw, unpublished data).

The cloned repeats were sequenced using the method of Sanger et al. (1977); part of an autoradiogram from a representative sequencing gel is shown in figure 2. The repeats were defined by having the Taq-I recognition sequence (TCGA) at both their 5' and 3' ends; therefore, we have included two of the four bases at the 5' (GA) and 3' (TC) ends as part of each repeat. Because the fragment is only 168 bp long, its sequence can be readily assessed from a single, 8% sequencing gel 0.75 m long. The pSEA1 sequence was determined by carrying out the sequencing reactions on four independent M13 clones, three of which contained the same strand and one the opposite strand. No ambiguous positions occurred in the sequence. In particular, the 168-bp sequence analyzed in this study appeared to be free of regions that cause the "pileups" of the type discussed by Bankier and Barrell (1983). The M13 clones from the population analyses were examined by replicate sequencing reactions, and once again no ambiguous positions occurred in the sequences. For the analysis of the IA individuals, replicate sequencing reactions were carried out for eight of the total of 39 clones examined. A consensus sequence was determined for each of the individuals analyzed by scoring the most common nucleotide at each position among the clones.

End Labeling and DNA Sequencing with Base-specific Chemical Cleavages

The 168-bp monomers of uncloned Taq-family repeats were isolated by TaqI digestion, agarose gel electrophoresis, and electroelution (see Cloning of the Highly Repetitive Sequence from Caledia) and end labeled by the "fill-in" reaction (Goodman 1979). These fragments were then digested with RsaI and subjected to the DNA sequencing reactions described by Maxam and Gilbert (1980).

Results

I. Highly Repetitive DNA in Caledia captiva

An initial characterization of highly repetitive DNA was carried out by examining the fraction recovered when total DNA was sheared to an average length of 500 bp, denatured, and then renatured to a Cot of 0.02; 1%–2% of the DNA from populations 2A and 4A of C. captiva was recovered in such a Cot 0.02 fraction. Buoyant density analyses in CsCl showed the presence of a hypersharp peak in both of these populations as well as a broad range of the other DNA species (fig. 3). The hypersharp peaks were of particular interest in this study because they indicated the presence of a major class of repetitive DNA that on renaturation could form high-molecular-weight aggregates, a characteristic of simple repetitive sequences. Furthermore, the fact that the hypersharp peaks from populations 2A and 4A were demonstrably different when the two hypersharp peaks were mixed and banded together (fig. 3) indicated that shifts in the base composition of the highly repetitive DNAs had occurred in the differentiation of races 2 and 4. Sequences from population 2A are the subject of this paper, whereas
sequences from population 4A are presently being characterized (M. L. Arnold and D. D. Shaw, unpublished data).

II. The Taq-Family Repetitive Sequence

Saturation hybridization experiments for population 1A from C. captiva indicated that approximately 0.2% of the total DNA, or $1.7 \times 10^5$ copies of the Taq-family sequence, were present in the haploid genome.
The chromosomal distribution of the \textit{Taq}I sequence was examined by means of in situ hybridization using a $^3$H c-RNA probe synthesized from the pSEA1 insert (the nucleotide sequence of this DNA is also shown in fig. 4). Numerous interstitial and
III. Sequence Variation in the Taq Family of Repeated Sequences
(a) Sequence Variation among Taq-Family Repeats from Two Population 1A Individuals

Individual copies of the 168-bp repeat were isolated from two population 1A individuals (1A-6 and 1A-19), and direct sequence comparisons were performed. When one examines the nucleotide sequences (fig. 5), numerous variant sites (58 variable positions among the 39 clones) that are spread throughout much of the sequence are immediately apparent. However, within this high level of variation there are stretches of unaltered sequence, the longest being 18 bp (position 67–84). This 18-bp conserved region is found in each of the 39 clones (fig. 5), is 72% A+T, and contains two direct repeats (CATT) separated by an adenine.

When each of the clones is compared to the consensus sequence determined for the individual from which they were derived (fig. 5), a wide range of variation becomes apparent. The cloned repeats from 1A-6 and 1A-19 differ from their consensus sequence by 1.0%–6.0% (1–10 bp) and 1.0%–9.0% (1–15 bp), respectively. Comparison of the two consensus sequences reveals different bases at positions 3, 6, 16, and 60, positions that correspond to the highly variable positions in the cloned repeats (fig. 5). Pairwise comparisons were carried out among the clones from each of the 1A individuals, and the results of these comparisons are given in table 1. From 0% to 10% variation in base-pair sequence is present within these two individuals, with a mean of 4% and 5% for 1A-19 and 1A-6, respectively.

(b) An 18-bp Sequence from within the Taq-Family Repeat Is Conserved in Caledia captiva

Figure 6 presents sequence data for clones from eight C. captiva populations representing the four races. Comparing each of the repeats shown in figure 6 (with the exception of the population 4A sequences) with the consensus sequence based on the 39 clones from 1A-6 and 1A-19 indicated that the divergence ranged from 1.0% to 7.0%. In addition, pairwise comparisons were made among all of the 11 clones from these races. From this analysis values of 6%, 5%, and 5% base-pair divergence were detected for the intrapopulational, interpopulational, and interracial comparisons, respectively (table 1). Much of this variability was located at the 58 positions previously identified in the 1A individuals. An additional 14 variable positions were, however, identified by the intraspecific comparisons. Thus, a total of 72 variable positions were identified among the 50 clones analyzed from populations 1A, 1C, 2A, 2B, 3A, 3B, and 3C. In spite of this high level of variability, the 18-bp region conserved in individuals 1A-6 and 1A-19 remained invariant; the probability of occurrence of this stretch departed significantly from that which would be expected to occur randomly ($P < 0.01$) when tested by the method suggested by Brown and Clegg (1983).

In addition to the analysis of the above clones, three repeat units were isolated and sequenced from population 4A. The copy number of the Taq-family repeat units in this population is only 2% of that found in population 1A (M. L. Arnold and D. D. Shaw, unpublished data). It is clear from inspecting these sequences (fig. 6) that Taq-family repeats in this population are highly heterogeneous. One repeat has only...
Fig. 5.—Repeats from two individuals—1A-6 (sequences 1–19) and 1A-19 (sequences 20–39)—and their respective consensus sequences (40 and 41). The 18-bp conserved region is indicated by boldface type. Sequences 1, 20, and 40 are written in their entirety. For the other sequences—2–19, 21–39, and 41 (the 1A-19 consensus sequence)—only those nucleotides that differ from the complete sequence immediately above them are shown. Boxes (□) indicate the absence of a base (A = adenine, C = cytosine, G = guanine, and T = thymine). Arrows (↑) located above the 1A-6 consensus sequence (40) denote positions at which substitutions, additions, or deletions occur among the 39 cloned repeats.
Table 1
Mean Base-Pair Change and Range of Base-Pair Change at Each of the Intraspecific Levels of Divergence

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Mean (Range) % of Base-Pair Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intraindividual</td>
<td>5.0 (0–10)</td>
</tr>
<tr>
<td>Intrapopulational</td>
<td>6.0 (4–8)</td>
</tr>
<tr>
<td>Interpopulational</td>
<td>5.0 (2–9)</td>
</tr>
<tr>
<td>Interracial</td>
<td>5.0 (1–10)</td>
</tr>
</tbody>
</table>

NOTE.—Mean base-pair divergence is the number of nucleotide positions that do not match (including gaps) times 100, divided by the length of the consensus sequence (i.e., 168 bp). The following pairwise comparisons were made: between clones from within individuals 1A-6 and 1A-19 (intraindividual values); between clones from within populations 2A, 2B, 3A, and 3C (intrapopulational values); between clones from populations within the same race (1A and 1C), (2A and 2B), and (3A, 3B, and 3C) (interpopulational values); and between clones from different races using populations 1A, 1C, 2A, 2B, and 3A, 3B, and 3C (interracial values).

four base-pair differences when compared to the consensus sequence from the population 1A individuals (fig. 6). (Subsequent experiments have suggested that this single clone may be a contaminant; this is presently under investigation.) This sequence also contains the 18-bp conserved region present in all of the preceding *C. captiva* repeats. In contrast, the two remaining sequences (Clones 13 and 14; fig. 6) are highly divergent when compared with the other *C. captiva* repeats (fig. 6). However, the 18-bp region discussed above is still recognizable even within these two divergent 4A sequences, although it is displaced 23 bp toward the 5' end of the repeat (fig. 6). In this region of the two divergent 4A clones, a total of 10 of the 18 bp are identical to the consensus sequence from the 1A individuals.

(c) The Consensus Sequence from 39 Clones of the Population 1A Repeat is Representative of the Taq-Family Repetitive Units

To relate the consensus sequence derived from cloned members of the *Taq* family to the entire population of *Taq*-family repeats, we utilized two different assays. One assay is based on the observation that an *RsaI* site is located 16 bp from the 5' end of the repeat unit and occurs in 56% of the clones from population 1A. To assay the frequency of this site in the entire population of *Taq*-family repeats, the DNA from approximately 40 individuals from population 1A was digested with *TaqI* and the band containing the 168-bp repeats was recovered from a preparative agarose gel. The DNA was end labeled using the fill-in reaction (see Material and Methods) and digested with *RsaI*. Electrophoresis of the digest on a 6% polyacrylamide gel resolved three fragments (168 bp, 152 bp, and 16 bp). Densitometric analysis of an autoradiogram of the gel indicated that 38% of the *Taq*-family repeats contained the *RsaI* site. The less frequent occurrence of the *RsaI* site would be predicted based on the assumption that the band of DNA that corresponded to the 168-bp repeats (see above) would also include other DNA species. Therefore, this frequency is consistent with the interpretation that the cloned sequences are not a highly selected subpopulation of repeat units.
Fig. 6.—Nucleotide sequences for cloned repeats from populations 1A(1), 1C(2), 2A(3, 4), 2B(5, 6), 3A(7, 8), 3B(9), 3C(10, 11), and 4A(12–14), along with the consensus sequences for races 1(15), 2(16), and 3(17). Boldface type indicates the 18-bp conserved region in repeats 1–14 and the consensus sequences (15–17). Sequences 1, 12, 13, and the race 1 consensus sequence, 15, are written in their entirety. The format and symbols are as given in fig. 5. Arrows (1) located above the race 1 consensus sequence (15) denote positions at which substitutions, additions, or deletions occur among either any of the sequences 1–11 in this figure or sequences 1–39 in fig. 5.
The second assay utilized the above 152-bp end-labeled RsaI fragment in a Maxam and Gilbert (1980) sequencing protocol (see Material and Methods). Preliminary sequence data for positions 72-116 (using the consensus sequence from population 1A as a reference) was obtained. This region was of particular interest because it included much of the highly conserved 18-bp region discussed above. Only three ambiguous positions (72, 77, and 83) were detected, indicating once again that the consensus sequence presented in figure 5 is a useful approximation of the consensus sequence for the population of Taq-family repeat units in the genome.

Discussion

This study has revealed extensive variation in nucleotide sequences at each of the levels of comparison, from intraindividual to interracial. Similar amounts of variation have been reported by Strachan et al. (1985) between cloned repeats of the 360-bp family isolated from pooled samples of *Drosophila* from a single species. In addition, a number of clones from several *Drosophila* species were compared, and in all but one of these comparisons there was a tenfold increase in the amount of divergence among the different cloned repeats (Strachan et al. 1985). This observation of extensive divergence between species with relatively low levels of within-species variation in both multigene families and highly repeated DNA has been termed concerted evolution (Zimmer et al. 1980; Dover 1982; Arneheim 1983). Although there have been numerous documentaries of this phenomenon between species, no study has extensively investigated the question of how variation is partitioned among populations that have not yet diverged to the level of biological species. The data presented in the *Caledia* study are directly relevant to the question of whether or not sequence variation in highly repeated DNA is correlated with levels of divergence below that of the species.

Shaw et al. (1980) and Coates and Shaw (1984) have described the patterns of reproductive isolation within and among the *C. captiva* races. The level of genetic compatibility between the different populations, in crosses involving different races among backcross and F2 generations, ranged from no reduction in viability to a 50% and 100% reduction in these hybrid types, respectively. Allozyme data for each of the four races of *C. captiva* have shown that protein electrophoretic divergence is correlated with the levels of reproductive isolation (Daly et al. 1981). Whatever mechanisms cause sequence divergence in the highly repeated *Taq* family, they are not directly correlated with genetic divergence below the species level, since the mean percentage of nucleotide divergence between repeats from a single individual is the same as that found in comparisons of races (table 1).

In contrast to the high levels of both intraspecific variation and interspecific divergence among highly repeated DNA families, numerous studies have detected significant similarities between tandemly repeated sequences from related animal species (see Singer 1982 for review). The estimates of sequence similarity were derived indirectly from melting-point determinations, positions of restriction-endonuclease cleavage sites, and the length of the basic repeating unit. For example, Strachan et al. (1982) used each of these criteria in their analysis of the evolution of several tandemly repeated DNA families from *Drosophila* species. Two repeated DNA families (360 bp and 500 bp) are highly similar in both *D. simulans* and *D. mauritiana*, whereas significant differences were observed between *D. yakuba* and *D. teissieri*. Lam and Carroll (1983a) have presented data on a 741-bp repeat from *Xenopus laevis* that, on the basis of restriction-endonuclease cleavage sites, indicate that variation among the repeats
from this family is extensive. No significant similarity was found between this sequence from *X. laevis* and the genomes of two other *Xenopus* species. In contrast, a 388-bp repeated family did share considerable similarity with sequences from two other *Xenopus* species (Lam and Carroll 1983b). Significant similarities between highly repeated sequences have also been detected for *Mus* species and between highly repeated sequences of both *Ovis* and *Bos*. Pietras et al. (1983) identified a minor satellite sequence in a *M. musculus* repetitive DNA library for which the repetition frequency was 5%–10% of that of the major satellite. Comparison of this minor satellite with the major satellite from *M. musculus* resolved a 29-bp region of extensive similarity between these two highly repeated families. Furthermore, this study demonstrated that sequences derived from *M. spretus* are also related to the minor satellite from *M. musculus*. Finally, a recent analysis of sheep highly repeated DNA has shown that three distinct repeat families (370 bp, 435 bp, and 800 bp) can be resolved by restriction-endonuclease digestion with *EcoRI* (Novak 1984). Each of these repeat families is related to each of the others on the basis of its nucleotide sequences, and a comparison of the 435-bp repeat with the bovine 1.715 satellite sequence revealed >50% identity between these two sequences (Novak 1984).

These observations of conserved highly repeated DNA sequences among both closely and more distantly related species has led to the suggestion that there may be functional constraints on the nucleotide composition of these sequences. It has, however, been pointed out by both Ohta and Dover (1983) and Miklos (1985) that this conservation does not necessarily imply that natural selection is acting or has acted on these sequences. The homogeneity of repeats between species can also be explained on the basis of recent origin, in which case an insufficient amount of time has elapsed for the accumulation of mutations, chance fixation of certain nucleotide sequences, and/or a mechanism of biased conversion that leads to the fixation of a variant sequence (Arnheim 1983; Ohta and Dover 1984).

An examination of the *C. captiva* Taq-family repeats reveals that the variation in nucleotide sequence, although encompassing approximately half of the base positions, is nonrandom. The obvious plasticity at some sites but not at others seems to indicate that there are limitations on which of the sites can change at high frequency. Particularly striking is the occurrence of an 18-bp region that is unchanged in all but two of the clones. These two clones isolated from population 4A are homologous to the 168-bp repeat both on the basis of filter hybridization (M. L. Arnold, unpublished data) and direct nucleotide examination, albeit at a much reduced level when compared with the other intraspecific values. This divergence has resulted in the shifting of the base sequence corresponding to the 18-bp conserved region. Whether this is a consequence of mutations that have produced a new *TaqI* site or of a rearrangement that has transposed this 18-bp region cannot be determined.

A similar pattern of conservation was detected by Dennis and Peacock (1984) when they sequenced the 180-bp repeat derived from the maize knob heterochromatin. A 27-bp sequence was found to be unvaried among the 19 clones isolated from maize, teosinte, and *Tripsacum*. The presence of a highly conserved region within a sequence that otherwise shows extensive variation suggests the presence of a mechanism that prevents the accumulation of mutations at these base-pair positions. One mechanism that could generate such a nonrandom pattern of change is biased gene conversion (Dover 1982) with the domain of conversion being less than the repeat length of the Taq-family sequence. One result of this process would be the generation of repeats with a mosaic pattern of change, with different regions of the sequence demonstrating
different amounts of variation, while the same region in different repeats would be tending toward homogeneity (Ohta and Dover 1984).

Clearly, many possibilities exist to explain the nonrandom pattern of divergence observed in \textit{C. captiva}, but it is interesting that the conserved 18-bp region within the Taq-family repeats represents part of a twofold symmetry (fig. 7) of the type found in DNA sequences that bind proteins (Ptashne et al. 1980; Siebenlist et al. 1980; Borgmeyer et al. 1984). Previously, Hsieh and Brutlag (1979) reported the presence of twofold symmetry in a \textit{Drosophila} heterochromatic sequence (the 360-bp family) that was shown to bind preferentially to an embryo-specific protein. In addition, highly repeated DNA (alphoid sequences) from primate species also demonstrates regions of dyad symmetry (see Miklos 1985 for review). However, in both of these examples there is no apparent sequence conservation in the areas of twofold symmetry when compared with other regions of the repeat. For example, the area of dyad symmetry in the 360-bp family from \textit{Drosophila} was found to be heterogeneous even among repeats from a single laboratory strain (Miklos and Gill 1981).

The confirmation of a function for the \textit{C. captiva} Taq family of repeats requires, first, the demonstration of a specific interaction of this sequence with a protein component, and second, the identification of the role of such an interaction. However, the conservation of a portion of this repeated element in conjunction with the unique structure of this conserved region is suggestive of such a cellular interaction.

Acknowledgments

We wish to thank Dr. David Coates for his assistance with field collecting and Frances Arnold for her typing of numerous drafts of the manuscript. We also thank Gary Brown and Maureen Whitaker for their assistance with the illustrations and photography, respectively. One of us (M.A.) is grateful to the Division of Plant Industry, CSIRO, for providing laboratory facilities for some of the experiments performed. M.A. was supported by an Australian National University Ph.D. scholarship. Excellent technical advice was given by P. Wilkinson, L. Moran, F. Clay, and B. Clarke. We are grateful to B. John and two anonymous reviewers for their helpful comments on the manuscript.

LITERATURE CITED


Wesley M. Brown, reviewing editor

Received January 15, 1985; revision received July 26, 1985.
Evolutionary History of the Hybridogenetic Hybrid Frog

*Rana esculenta* as Deduced from mtDNA Analyses

Christina Spolsky² and Thomas Uzzell²

Academy of Natural Sciences of Philadelphia

mtDNA of the hybridogenetic hybrid frog *Rana esculenta* from Switzerland, Austria, and Poland was compared to mtDNA of the parental species *R. ridibunda* and *R. lessonae* using electrophoretic analysis of restriction enzyme fragments. Two mtDNA phenotypes, with 3.4% sequence divergence, are present in *R. lessonae*: type C is found in Poland, and type D is found in Switzerland. *Rana ridibunda* from Poland has either of two mtDNA phenotypes: type A is the typical ridibunda mtDNA, and type B is a lessonae mitochondrial genome, introgressed into *R. ridibunda*. that differs from type C mtDNA of *R. lessonae* by only 0.3%. Each of the three lessonae genomes differs from A, the typical ridibunda mtDNA, by ~8%. All four types of mtDNA (A and B of *R. ridibunda*, C and D of *R. lessonae*) are found in *R. esculenta*. Of 62 *R. esculenta* from Poland, 58 had type C, three had type A, and one had type B mtDNA. All nine *R. esculenta* from Switzerland had type D mtDNA. All three *R. esculenta* from Austria, from a population in which males of *R. esculenta* are rare, had ridibunda mtDNA, two having type B and one having type A. Both field observations and studies of mating preference indicate that the primary hybridizations that produce *R. esculenta* are between *R. ridibunda* females and *R. lessonae* males; thereafter, *R. esculenta* lineages are usually maintained by matings of *R. esculenta* females with *R. lessonae* males. The presence of ridibunda mtDNA in the three *R. esculenta* sampled from Austria, its occasional presence in *R. esculenta* populations in Poland, and its absence from *R. esculenta* in Switzerland support both the direction of the original hybridization and the rarity of formation of new *R. esculenta* lineages. The preponderance of *R. esculenta* individuals with lessonae mtDNA in our samples from central Europe suggests that most lineages have gone through at least one mating between an *R. lessonae* female and an *R. esculenta* male. This reveals a greater reproductive role for *R. esculenta* males than their partial sterility and infrequent matings would suggest.

Introduction

Among the 50-odd clonally reproducing species of fishes, amphibians, and reptiles, *Rana esculenta* Linnaeus 1758, the common edible frog of Europe, is unique in that, in most populations, males as well as females occur in large numbers. Even so, the reduced fertility of males (Berger 1970, 1971; Günther 1973), the relative inviability of progeny fathered by them (Berger and Uzzell 1977), and their tendency to engage in combative rather than reproductive behavior in breeding congregations (Blankenhorn 1974, 1977) suggest that they have relatively little to do with maintaining populations of *R. esculenta*.

1. Key words: mtDNA, hybridogenesis, *Rana esculenta*, evolutionary pathways.


Address for correspondence and reprints: Dr. Christina Spolsky, Museum of Natural History and Department of Ecology, Ethology and Evolution, University of Illinois, Urbana, Illinois 61801.

© 1986 by The University of Chicago. All rights reserved.
0737-4038/86/0301-3105$02.00
Rana esculenta arose and, in areas of sympatry, arises by hybridization between the two Mendelian species R. lessonae Camerano 1882 and R. ridibunda Pallas 1771 (Berger 1967, 1968). Rana ridibunda is a large frog, distinguishable from R. lessonae on the basis of body size, call, and numerous morphological and biochemical features. Rana lessonae is a much smaller, more terrestrial, frog; the hybrid R. esculenta is intermediate in size and other morphological features between the two parental species. Of particular interest are the size differences between the taxa. The size range for sexually mature R. ridibunda (74–94 mm) does not overlap that for R. lessonae (42–71 mm); Rana esculenta spans the range (54–89 mm) between the two parental species (Berger 1966, 1970).

Both behavioral studies and field observations indicate that the original hybridizations that produce R. esculenta are between female R. ridibunda and male R. lessonae. Such matings—but never the reverse—between these two species have been observed in the field on many separate occasions (Berger 1957, 1959, 1970; Borkin et al. 1979). This directionality of interspecific matings depends both on size preferences in mating shown by males and on the large size difference between R. ridibunda males and R. lessonae females. In laboratory studies (Blankenhorn 1974, 1977), male water frogs showed a marked preference for larger females. In various combinations of pairings of R. esculenta and R. lessonae, larger females were preferred over smaller ones. No sexual behavior was displayed in any combination in which the female was smaller than the male; in fact, males actively avoided females smaller than themselves. As in most species of frogs (Wright and Wright 1949, p. 20), male water frogs are smaller on the average than conspecific females (Berger 1966, Borkin et al. 1978). Since the smallest sexually mature males of R. ridibunda are almost invariably longer than the largest female R. lessonae, the size difference, coupled with the size preference in mating shown by water frog males, virtually precludes pairing of R. ridibunda males and R. lessonae females in nature (Tunner 1974; H. Hotz, personal communication).

Reproduction in both sexes of R. esculenta is hybridogenetic (Tunner 1973). In hybridogenesis (Schultz 1969), the genome of one parental species is eliminated from germ line cells before the completion of gametogenesis but normally is restored at fertilization because the hybridogenetic individuals mate with that parental species. In the common form of R. esculenta in central Europe, the lessonae chromosome set is absent from both sperm and ova (Tunner 1974; Graf and Müller 1979; Uzzell et al. 1980), which contain only a ridibunda chromosome set. Such R. esculenta live with and reproductively depend on R. lessonae (the L-E system; Uzzell and Berger 1975). In each L-E system lineage, the ridibunda nuclear genome is passed clonally from generation to generation, while a new lessonae genome enters at each fertilization only to be lost before the next fertilization (fig. 1). In the L-E system, R. esculenta lineages are maintained mostly by matings between R. esculenta females and R. lessonae males (Blankenhorn 1974, 1977; L. Berger, personal communication), but because of the overlap in sizes, the reverse mating pattern—that is, between a large Rana lessonae female and a small Rana esculenta male—does occur at low frequency, both in nature and in the lab (Blankenhorn 1974; L. Berger, personal communication).

As a key to exploring the population dynamics and the evolutionary history of these frogs, we have examined mitochondrial DNA from numerous R. esculenta of the L-E system as well as from the two parental species. The maternal inheritance of mtDNA in Metazoa (Dawid and Blackler 1972; Hutchinson et al. 1974; Kroon et al. 1978; Giles et al. 1980) makes this genome useful in tracing maternal genealogies (Avise et al. 1979a, 1979b; Brown and Wright 1979; Ferris et al. 1981b; Brown et al.
**FIG. 1.**—An overview of mating patterns within the *Rana esculenta* complex. Large boldface letters indicate the chromosomal complement of diploid individuals; single capital letters with dotted-line arrows to them indicate meiotic products. The mitochondrial types in diploid individuals and in ova are designated by shadings—stippled for ridibunda mtDNA and cross-hatched for lessonae mtDNA. Thin-line arrows represent pairings in which the male is to the right of the female and hence likely to be larger—and in which, therefore, the mating is less frequent than propinquity would allow. Thick-line arrows represent the reverse situation, in which the male is likely to be smaller than the female (the favorable case for mating). The LL female × RR male mating presumably does not occur. All other possibilities not shown are either intraspecific or are rare and produce *R. ridibunda*. Mating 1 produces *R. esculenta*; 2 maintains *R. esculenta* with ridibunda mitochondria; 3 introduces lessonae mitochondria into *R. esculenta*; 4 maintains *R. esculenta* with lessonae mitochondria; and 5 transfers lessonae mitochondria on to *R. ridibunda*.

1982; Hauswirth and Laipis 1982; Wright et al. 1983). In the present study, maternal genealogies traced by means of restriction fragment analysis of mtDNA from *R. esculenta* and from its parental species, *R. ridibunda* and *R. lessonae*, reveal a greater role in reproduction for male *R. esculenta* than previously has been suspected.

**Material and Methods**

**Specimens**

Frogs were identified as to species by morphology and electrophoretic phenotype. Ploidy was determined using red blood cell sizes (Berger and Uzzell 1975; Günther 1977); all *Rana esculenta* sampled were diploid. Most samples of frogs of all three species (132 individuals) were collected from six localities in western Poland, within
a 40-km radius of Poznan. At some of these localities only *R. lessonae* and *R. esculenta* occurred, but all three species occurred at others. These six localities were grouped in the analysis because no significant differences between localities in mtDNA ratios were found. In addition, nine *R. esculenta* were obtained from two localities near Zürich, Switzerland; eight *R. lessonae* came from Fehraltdorf, Canton Zürich, Switzerland; and three *R. esculenta* were collected at Neusiedlersee, Austria. All *R. ridibunda* came from Poland (table 1).

Preparation of mtDNA

mtDNA was isolated from individual frogs using methods described by Spolsky and Uzzell (1984). In some cases the mitochondrial fraction was enriched by banding in a sucrose step gradient (0.9 M and 1.8 M sucrose) prior to lysis. Purified mtDNAs were redissolved in 50–300 μl of 0.1 TE (1 mM tris, 0.1 mM ethylenediaminetetraacetate, pH 8.0) and stored at -70°C. The amount of mtDNA was estimated by minigel electrophoresis and ethidium bromide staining of a 5-μl aliquot of each preparation.

Restriction Endonuclease Analysis of mtDNA

Approximately 5–10 ng of each DNA were digested to completion with each restriction enzyme (table 2) under conditions recommended by the supplier (Boehringer Mannheim Biochemicals or Bethesda Research Labs). Resulting DNA fragments were end-labeled with a mixture of four (adenine, cytosine, guanine, and tyrosine) α-32P-triphosphate deoxynucleosides and subjected to electrophoresis through 1% agarose gels and, for small fragments, 4% polyacrylamide gels (Brown 1980; Wright et al. 1983); separated fragments were detected by autoradiography. For each gel, fragment sizes were estimated from mobilities of DNA fragments of known size (HindIII-restricted lambda and PM2 DNAs, HincII-restricted φX174 DNA).

Estimation of Sequence Divergence and mtDNA Relatedness

The amount of sequence divergence was calculated from the fraction of restriction fragments shared by a pair of DNAs (Nei and Li 1979), using the formula derived by Upholt (1977). An unrooted Wagner network of mtDNA relationships was constructed from the matrix of divergence values.

Results

mtDNAs in the Parental Species

Four types of mtDNAs were found in *Rana ridibunda* and *R. lessonae*. For comparison between species and populations, fragment patterns were determined for one individual of each mtDNA type using 19 (types A, B, and C; Spolsky and Uzzell 1984) or 15 (type D; table 2) restriction enzymes. Paired comparisons of sequence differences were made for the four types of mtDNAs found in *R. ridibunda* and *R. lessonae* (table 3). Type A mtDNA of *R. ridibunda* is most divergent from the other three types; the sequence divergence of A from any of the other three types is approximately the same, 8.1%–8.5%. Types B and C are virtually identical, with a difference in sequence of 0.3%, so their mean divergence from A and their mean divergence from D were used for the analyses. The sequence difference between the D and C mtDNAs of *R. lessonae*, 3.4%, is within the range of intraspecific differences found in other vertebrates (Avise et al. 1979b; Ferris et al. 1981a; Wolstenholme et al. 1982; Lansman et al. 1983).
Table 1
Restriction Fragment Patterns in Type C and D mtDNAs*

<table>
<thead>
<tr>
<th>ENZYME</th>
<th>D</th>
<th>C*</th>
<th>ENZYME</th>
<th>D</th>
<th>C*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Avi</em></td>
<td>8,800</td>
<td>7,000</td>
<td><em>HindIII</em></td>
<td>5,700</td>
<td>5,700</td>
</tr>
<tr>
<td></td>
<td>7,000</td>
<td>6,000</td>
<td></td>
<td>4,000</td>
<td>4,000</td>
</tr>
<tr>
<td></td>
<td>2,950</td>
<td>5,100</td>
<td></td>
<td>4,000</td>
<td>4,000</td>
</tr>
<tr>
<td></td>
<td>850</td>
<td>1,100</td>
<td></td>
<td>2,150</td>
<td>2,150</td>
</tr>
<tr>
<td></td>
<td>130b</td>
<td>130b</td>
<td></td>
<td>1,500</td>
<td>1,500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>105</td>
<td></td>
<td>1,100</td>
<td>1,100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>600b</td>
<td>600b</td>
</tr>
<tr>
<td><em>BamHI</em></td>
<td>13,000</td>
<td>11,000</td>
<td></td>
<td>10,500b</td>
<td>10,500b</td>
</tr>
<tr>
<td></td>
<td>6,300</td>
<td>8,000</td>
<td></td>
<td>6,000b</td>
<td>6,000b</td>
</tr>
<tr>
<td></td>
<td>550</td>
<td>470</td>
<td><em>HpaI</em></td>
<td>3,250b</td>
<td>3,250b</td>
</tr>
<tr>
<td><em>BclI</em></td>
<td>17,000</td>
<td>8,400</td>
<td></td>
<td>4,900b</td>
<td>4,900b</td>
</tr>
<tr>
<td></td>
<td>2,150b</td>
<td>7,800</td>
<td></td>
<td>4,200b</td>
<td>4,200b</td>
</tr>
<tr>
<td></td>
<td>900b</td>
<td>2,150b</td>
<td><em>EcoRI</em></td>
<td>625b</td>
<td>625b</td>
</tr>
<tr>
<td></td>
<td></td>
<td>900b</td>
<td></td>
<td>500b</td>
<td>500b</td>
</tr>
<tr>
<td><em>BglII</em></td>
<td>5,300</td>
<td>5,700</td>
<td><em>EcoRV</em></td>
<td>19,500</td>
<td>16,000</td>
</tr>
<tr>
<td></td>
<td>3,750b</td>
<td>3,750b</td>
<td></td>
<td>3,500</td>
<td>3,500</td>
</tr>
<tr>
<td></td>
<td>3,600b</td>
<td>3,600b</td>
<td></td>
<td>3,500</td>
<td>3,500</td>
</tr>
<tr>
<td></td>
<td>1,650b</td>
<td>3,000b</td>
<td></td>
<td>4,900b</td>
<td>4,900b</td>
</tr>
<tr>
<td></td>
<td>1,400</td>
<td>1,650b</td>
<td><em>HaeII</em></td>
<td>8,500b</td>
<td>8,500b</td>
</tr>
<tr>
<td></td>
<td>1,200</td>
<td>1,500</td>
<td></td>
<td>4,200b</td>
<td>4,200b</td>
</tr>
<tr>
<td></td>
<td>1,100</td>
<td>760b</td>
<td></td>
<td>625b</td>
<td>625b</td>
</tr>
<tr>
<td></td>
<td>760b</td>
<td>220</td>
<td><em>PvuII</em></td>
<td>14,200</td>
<td>15,000</td>
</tr>
<tr>
<td><em>HincII</em></td>
<td>5,200</td>
<td>5,600</td>
<td></td>
<td>4,300</td>
<td>2,250</td>
</tr>
<tr>
<td></td>
<td>3,200b</td>
<td>3,300</td>
<td></td>
<td>600</td>
<td>270</td>
</tr>
<tr>
<td></td>
<td>2,900</td>
<td>3,200b</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2,150b</td>
<td>1,900</td>
<td><em>KpnI</em></td>
<td>7,300b</td>
<td>7,300b</td>
</tr>
<tr>
<td></td>
<td>1,700b</td>
<td>1,700b</td>
<td></td>
<td>7,300b</td>
<td>7,300b</td>
</tr>
<tr>
<td></td>
<td>1,300b</td>
<td>1,300b</td>
<td></td>
<td>3,300b</td>
<td>3,300b</td>
</tr>
<tr>
<td></td>
<td>1,230</td>
<td>1,050</td>
<td></td>
<td>1,300b</td>
<td>1,300b</td>
</tr>
<tr>
<td></td>
<td>1,150</td>
<td>830</td>
<td></td>
<td>620b</td>
<td>620b</td>
</tr>
<tr>
<td></td>
<td>730</td>
<td>340b</td>
<td><em>PstI</em></td>
<td>12,500b</td>
<td>12,500b</td>
</tr>
<tr>
<td></td>
<td>340b</td>
<td>110</td>
<td></td>
<td>3,600</td>
<td>7,100</td>
</tr>
<tr>
<td></td>
<td>290</td>
<td></td>
<td></td>
<td>3,600</td>
<td>3,600</td>
</tr>
<tr>
<td><em>SmaI</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19,500</td>
</tr>
<tr>
<td><em>XbaI</em></td>
<td>9,800</td>
<td></td>
<td></td>
<td></td>
<td>19,500</td>
</tr>
<tr>
<td></td>
<td>9,800</td>
<td></td>
<td></td>
<td></td>
<td>9,800</td>
</tr>
</tbody>
</table>

* Fragment patterns for type C mtDNA, as well as for types A and B mtDNA, have been previously published (Spolsky and Uzzell (1984)).

b Shared fragments.

c The difference in the next-to-smallest KpnI fragment represents size variation; we consider these fragments homologous in mtDNA types C and D.

These sequence divergence values were used to construct a Wagner network (fig. 2). Most of the divergence in the network, 6.6, is in the leg between A and the common node. The distances from D and from the B-C pair to the common node
Table 2
Distribution of mtDNA Types in the Rana esculenta Complex

<table>
<thead>
<tr>
<th>SPECIES AND REGION (N)</th>
<th>mtDNA TYPE (Frequency)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>R. ridibunda: Poland (38)</td>
<td>...</td>
</tr>
<tr>
<td>R. lessonae: Poland (32)</td>
<td>...</td>
</tr>
<tr>
<td>Switzerland (8)</td>
<td>...</td>
</tr>
<tr>
<td>R. esculenta: Poland (62)</td>
<td>...</td>
</tr>
<tr>
<td>Switzerland (9)</td>
<td>...</td>
</tr>
<tr>
<td>Austria (3)</td>
<td>.33</td>
</tr>
</tbody>
</table>

are approximately the same, being 1.9 and 1.5, respectively. If the network is rooted at the midpoint of the longest inter-mtDNA distance, types B, C, and D cluster together to form a lessonae-like group.

Because BamHI and SmaI are the only two restriction enzymes that distinguish between B and C mtDNAs and because they also distinguish among all four mtDNA types (fig. 3), these two enzymes were used to classify each mtDNA sample into one of the four types.

BamHI cuts A mtDNA at a single position to produce a full-length linear molecule; it cleaves B mtDNA into three fragments approximately 8,000, 6,400, and 5,600 bp in length; and it cleaves C mtDNA into three fragments of 11,500, 8,000, and 470 bp. Type D mtDNA is cut at three sites to generate fragments 13,000, 6,300, and 550 bp in length (fig. 3).

SmaI digestion produced no common fragments between any of the four types. An mtDNA is cut by SmaI at three sites into fragments approximately 8,800, 8,800 and 1,800 bp long. B is cleaved into two fragments 14,000 and 5,500 bp long, and C is cut at a single site. There are no SmaI restriction sites on D mtDNA (fig. 3).

In addition to BamHI and SmaI, between one and 10 other restriction enzymes were used on ~25% of the samples. Although more extensive sampling would probably have demonstrated some variation, in our limited sampling no gain or loss of restriction

Table 3
Matrix of mtDNA Comparisons

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>...</td>
<td>.256</td>
<td>.258</td>
<td>.244</td>
</tr>
<tr>
<td>B</td>
<td>8.2 ± 1.0</td>
<td>...</td>
<td>.946</td>
<td>.556</td>
</tr>
<tr>
<td>C</td>
<td>8.1 ± 1.0</td>
<td>.3 ± .3</td>
<td>...</td>
<td>.560</td>
</tr>
<tr>
<td>D</td>
<td>8.5 ± 1.1</td>
<td>3.4 ± .8</td>
<td>3.4 ± .8</td>
<td>...</td>
</tr>
</tbody>
</table>

NOTE.—Values above the diagonal are the proportion of shared fragments; values below the diagonal are the percentages of sequence difference ± SD.
sites was observed within any of the four mtDNA types; variation in length of mtDNA within types was detected, however.

Individuals of *R. ridibunda* from Poland had one or the other of two very different types of mtDNA, i.e., A or B (Spolsky and Uzzell 1984; table 1). With the restriction enzymes used, mtDNA of *R. lessonae* was locally invariant but differed between Switzerland and Poland. *Rana lessonae* from Poland had type C mtDNA, whereas in Switzerland this species had type D mtDNA (table 1).

**mtDNA in *R. esculenta***

In screening *R. esculenta* populations, each mtDNA sample was identified as to type by restriction with *BamHI* and *SmaI*. Most *R. esculenta* from central Europe that were examined had *R. lessonae* mtDNA, rather than either of the mtDNA patterns found in *R. ridibunda* (table 1). All nine *R. esculenta* from Switzerland had the

---

**Fig. 2.**—Inheritance patterns in hybridogenetic L-E system *Rana esculenta*. *Rana esculenta* hybrids originate from matings of *R. ridibunda* females with *R. lessonae* males. In the L-E system, diploid *R. esculenta* passes on only the ridibunda chromosome set through its gametes. *Rana esculenta* hybrid lineages are usually maintained by crosses of *R. esculenta* females with *R. lessonae* males, although the reverse cross—*R. lessonae* female × *R. esculenta* male—does rarely occur.

---

**Fig. 3.**—Rootless Wagner network of mtDNA relatedness. Distances are percentages of sequence divergences. CH = Switzerland; PL = Poland. The *Rana ridibunda* samples came from Poland. A, B, C, and D are the four mitochondrial types found.
D type of mtDNA found in *R. lessonae* in Switzerland. Of the 62 *R. esculenta* from Poland, 58 had the same mtDNA type, C, as did *R. lessonae* from this area. The remaining four *R. esculenta* from Poland had one of the mtDNA types seen in *R. ridibunda*: three, each from a different locality, had the A genome, and one had the B genome. A different pattern, however, was found in Austria: all three *R. esculenta* from there had ridibunda rather than lessonae mtDNA; two had B and one A mtDNA.

Discussion

Relationships within the *Rana esculenta* Complex

The distinction between *R. lessonae* and *R. ridibunda*, first clearly stated by Berger (1957), has been abundantly confirmed by subsequent morphological (Berger 1966), electrophoretic (Tunner 1970, 1972, 1973; Engelmann 1972, 1973; Uzzell and Berger 1975) and immunological (Uzzell 1979, 1982) studies. The mtDNA network (fig. 2), when rooted on the longest branch, suggests a phylogeny for *R. ridibunda* and *R. Zessonae* that is inconsistent with these species limits, because in the rooted network the type B genome found in *R. ridibunda* clusters with the lessonae genomes. The mtDNA phylogeny can be reconciled with the species limits by postulating an introgression of *R. lessonae* mtDNA into *R. ridibunda* during the relatively recent past (Spolsky and Uzzell 1984).

Immunological comparisons of serum albumins place the divergence time of *R. ridibunda* and *R. lessonae* at ~12 Myr ago (Uzzell 1982). With a sequence divergence between the two species for A and C mtDNAs of 8.5% ± 1.1% (table 3), the rate of sequence change for mtDNA in this species pair is ~0.7%/Myr. If this rate of change were constant over this time period, which is far from certain, it would appear that the introgressed B mtDNA of *R. ridibunda* diverged from type C mtDNA of *R. lessonae* within the last 400,000 years and that mitochondrial genomes assayed from Swiss and Polish *R. lessonae* last shared a common ancestor 4–6 Myr ago. Since *R. lessonae* reinvaded the entire region between Switzerland and Poland after the Würm glaciation, probably within the past 15,000 years, the divergence time for the mtDNAs suggests that northern Europe was repopulated by already divergent stocks (cf. Avise et al. 1984).

Evolutionary Pathways in *R. esculenta*

Since the primary hybridizations that produce *R. esculenta* lineages involve at least primarily, if not exclusively, crosses of *R. ridibunda* females with *R. lessonae* males and since L-E system lineages are usually maintained by matings of *R. esculenta* females with *R. lessonae* males, we expected to find ridibunda mtDNA in our samples of *R. esculenta*. Instead, the majority of *R. esculenta* individuals had lessonae mtDNA rather than either of the mtDNAs found in *R. ridibunda*. Two alternative hypotheses could explain this observation. Either (1) the original hybridizations were, contrary to expectations, mainly between *R. lessonae* females and *R. ridibunda* males or (2) each *R. esculenta* lineage has gone through at least one mating between an *R. esculenta* male and an *R. lessonae* female.

Both field observations and laboratory studies, as reviewed in the Introduction, make the first hypothesis unlikely (the presence of both ridibunda and lessonae mtDNA in *R. esculenta* would in any case require a few original hybridizations involving *R. ridibunda* females). Furthermore, since lessonae mtDNA is very common in *R. es-
culenta, despite the fact that the few observed matings between the parental species all involve *R. ridibunda* females, some additional factor must be invoked to account for this distribution if the first hypothesis is correct.

Support for the origin of *R. esculenta* from crosses of female *R. ridibunda* with male *R. lessonae* is provided by the small sample of *R. esculenta* from Neusiedlersee in eastern Austria. The *R. esculenta* population there, as throughout the Pannonian Basin, is exceptional among L-E populations in that it consists almost entirely of females (Tunner and Dobrowsky 1976; Berger et al. 1985). Since male *R. esculenta* are very rare here, matings of male *R. esculenta* with female *R. lessonae* virtually never occur. Although one might expect, a priori, to find the same proportion (94%-100%) of *R. esculenta* with lessonae mtDNA in Austria as in the other localities, all three individuals of *R. esculenta* from Neusiedlersee have ridibunda mtDNA. The probability of picking three individuals at random with ridibunda mtDNA, if 95% of the population had lessonae mtDNA, is vanishingly small ($1.3 \times 10^{-4}$). Finding only ridibunda mtDNA in this population, where they can only very rarely be replaced with lessonae mtDNA, strongly suggests that the original hybridizations were between *R. ridibunda* females and *R. lessonae* males.

The alternative hypothesis that would account for the distribution of mtDNA in *R. esculenta*—that is, that males of *R. esculenta* have a significant role in maintaining populations of *R. esculenta* in the L-E system—is thus more plausible. It requires only one mating between a female *R. lessonae* and a male *R. esculenta* in each lineage. Once such a cross occurs in an *R. esculenta* lineage, that lineage becomes fixed for lessonae mtDNA, since subsequent crosses of the usual type (*R. esculenta* female × *R. lessonae* male) will maintain lessonae mtDNA in the lineage (fig. 4).

Given the small proportion of observed matings of esculenta males with lessonae females compared to the reverse cross, the widespread occurrence of lessonae mtDNA in *R. esculenta* is surprising. There are, however, a number of factors that could be responsible for the observed distribution: (1) Crosses between *R. lessonae* females and *R. esculenta* males may be more common than reported, although that alone would not account for the high proportion of *R. esculenta* with lessonae mtDNA. (2) There may be some selective advantage to *R. esculenta* in having lessonae mtDNA either in gametes or in the soma, although there is no direct evidence for this. (3) Sufficient time has elapsed for virtually all lineages to have gone through at least one mating of the rarer type; since a change to lessonae mtDNA is irreversible in an L-E system population (fig. 4), an increasing proportion of individuals with lessonae mtDNA will accumulate in a population as a result of additional matings of the rarer type in previously unchanged lineages. This is formally analogous to the increase in frequency of mutants in a bacterial system with a constant mutation rate (Kaplan 1947; Novick and Szilard 1950); it requires no assumption that either type of mtDNA in *R. esculenta* has a selective advantage over the other or that *R. esculenta* possessing either type of mtDNA reproduces more successfully. Crosses of the rarer type could thus account for most, if not all, of the distribution of lessonae mtDNA in *R. esculenta* in central Europe.

The increase with time of lessonae mtDNA in *R. esculenta* populations must be weighed against the increase of ridibunda mtDNA through the formation of new *R. esculenta* lineages. Since individuals of *R. esculenta* with ridibunda mtDNA may represent either old lineages that have not gone through a hybridogenetic mating of the unusual kind or newly arisen lineages, those few *R. esculenta* in Poland with ridibunda
mtDNAs also reflect the relative rarity of formation of new hybrid lineages. In Switzerland, where there are no autochthonous *R. ridibunda*, no primary hybridizations can occur; the lack of ridibunda mtDNA in *R. esculenta* in Switzerland suggests that all extant *R. esculenta* lineages there have gone through the rarer type of hybridogenetic mating.

To summarize, in areas where no primary hybridization can presently occur (e.g., most of Switzerland), *R. esculenta* do not have ridibunda mtDNA; where primary hybridizations can occur (e.g., Poland, Austria), at least some *R. esculenta* have ridibunda mtDNA; in Austria, where matings of male *R. esculenta* and female *R. lessonae* are extremely rare, *R. esculenta* have preponderantly, if not exclusively, ridibunda mtDNA. These data, taken together with behavioral data, fit the hypothesis that primary hybridizations are predominantly, if not exclusively, between female *R. ridibunda* and male *R. lessonae*; lessonae mtDNA gets into *R. esculenta* lineages through rare matings.
between female *R. lessonae* and male *R. esculenta*. The alternative hypothesis—that is, that most primary hybridizations are between female *R. lessonae* and male *R. ridibunda*—besides being contrary to behavioral data and field observations, cannot hold in Austria and does not easily explain the differences between mtDNA type distributions in *R. esculenta* in Switzerland and Poland. The pervasive presence of *lessonae* mtDNA in central European *R. esculenta* indicates, therefore, that *R. esculenta* males play a significant role in maintaining these hybridogenetic populations in Poland and Switzerland.

**Acknowledgments**

Leszek Berger, Hansjürg Hotz, and Heinz Tunner provided many of the animals. Janice Forsyth and Branin Boyd provided technical assistance. Maciej Pietczak (Institute of Plant Genetics, Polish Academy of Sciences, Poznan, Poland) and Kazimierz Ziemnicki (Physiology Department, Adam Mickiewicz University in Poznan) kindly provided laboratory facilities. This work was supported by grants from the National Science Foundation (DEB 81-11397) and the Smithsonian Institution’s Foreign Currency Program (T.U.) and from the Whitehall Foundation (C.S. and T.U.).

**LITERATURE CITED**


WESLEY M. BROWN, reviewing editor

Received July 23, 1985; revision received September 18, 1985.
The nucleotide sequences of four genes of the influenza A virus (nonstructural protein, matrix protein, and a few subtypes of hemagglutinin and neuraminidase) are compiled for a large number of strains isolated from various locations and years, and the evolutionary relationship of the sequences is investigated. It is shown that all of these genes or subtypes are highly polymorphic and that the polymorphic sequences (alleles) are subject to rapid turnover in the population, their average age being much less than that of higher organisms. Phylogenetic analysis suggests that most polymorphic sequences within a subtype or a gene appeared during the last 80 years and that the divergence among the subtypes of hemagglutinin genes might have occurred during the last 300 years. The high degree of polymorphism in this RNA virus is caused by an extremely high rate of mutation, estimated to be 0.01/nucleotide site/year. Despite the high rate of mutation, most influenza virus genes are apparently subject to purifying selection, and the rate of nucleotide substitution is substantially lower than the mutation rate. There is considerable variation in the substitution rate among different genes, and the rate seems to be lower in nonhuman viral strains than in human strains. This difference might be responsible for the so-called freezing effect in some viral strains.

Introduction

The influenza virus is a single-stranded RNA virus and is classified into three types, A, B, and C, according to immunological differences (Webster et al. 1982). The type A virus is the major cause of influenza epidemics not only in humans but in other mammals and birds. As do other RNA viruses, this virus has a high mutation rate, the mutation rate per year being approximately one million times higher than that of DNA organisms (Air 1981; Holland et al. 1982). Because of this high mutation rate, the influenza A virus becomes progressively resistant to antibodies made against older viruses. The high mutation rate is also responsible for the high degree of polymorphism observed in the genes in this organism (Nei 1983).

In recent years, the nucleotide sequences of several influenza virus genes have been determined for various strains, including those that have been kept in refrigerators for many years. These data provide a unique opportunity for studying the evolutionary history of polymorphic genes as well as for estimating the rate of nucleotide substitution in evolution.

Previously, Krystal et al. (1983) and Martinez et al. (1983) estimated the rate of nucleotide substitution for some of the genes of this virus, but their estimates are not very reliable because they simply compared genes from two strains isolated in different years. Nei (1983) also estimated the substitution rate by fitting a regression equation.

1. Key words: influenza A virus, phylogenetic tree, polymorphism, mutation rate, rate of nucleotide substitution.
to data from various strains isolated in different years. In the presence of polymorphism, however, his method is expected to give an overestimate. A better method of estimating the rate of nucleotide substitution for this case is first to examine the evolutionary relationship of genes obtained from different strains and then use only strains that are closely related by descent.

The main purpose of this paper is to conduct phylogenetic analyses of nucleotide sequences obtained from various strains and to estimate the rate of nucleotide substitution by using the method mentioned above.

**Nucleotide Sequences Used**

The genome of the influenza A virus consists of eight RNA segments coding for 10 different proteins (e.g., Lamb 1983). In the present study, we used four different genes, i.e., the hemagglutinin gene on segment 4, the neuraminidase gene on segment 6, the matrix protein 1 gene on segment 7, and the nonstructural protein 1 gene on segment 8. The first two genes are known to be responsible for the antigenic variation of this virus. We used only the coding regions of these genes, excluding the initiation codon. In the regions studied, there were no deletions or insertions. We used 46 strains in this study, and they are presented in table 1. Each strain is designated by the abbreviation of the location and the year in which it was isolated (see the legend to table 1).

The hemagglutinin gene is highly variable and can be classified into 13 subtypes (H1–H13) according to the immunological differences in hemagglutinin (Webster et al. 1982). The average nucleotide difference among these subtypes is 51% per nucleotide site (Nei 1983). We have therefore treated these subtypes separately, as though they were different genes. Sequence data useful for our analysis were available only for four subtypes, i.e., H1, H2, H3, and H11. The viral strains used for these four subtypes are given in table 1. The hemagglutinin gene is composed of three functional regions, i.e., signal peptide (SP), hemagglutinin 1 (HA1), and hemagglutinin 2 (HA2), and sequence data for these regions were combined unless otherwise mentioned. The nucleotide sequences available were not always complete, and the number of nucleotides used are presented in table 1.

The neuraminidase gene is also highly variable and can be divided into nine subtypes (NI–N9) (Webster et al. 1982). Two of them (NI and N2) were used here. Since sequence data for the NI subtype were available only for the first 168 bases in most strains, our analysis was restricted to this region. There were complete nucleotide sequences (1,404 bases) available for five N2 subtype strains. The coding region for the matrix protein 1 gene (MX1) overlaps with that for the matrix protein 2 gene (MX2) on segment 7, and, similarly, the coding region for the nonstructural protein genes 1 (NS1) and 2 (NS2) overlap with each other. Therefore, we used only the genes for MX1 and NS1. For gene NS1, there were two sets of data; partial sequences and complete sequences (see table 1). In the construction of phylogenetic trees, all sequence data were used, excluding unshared nucleotides. In the estimation of the rate of nucleotide substitution, however, only complete sequences were used.

**Phylogenetic Trees**

Phylogenetic trees were constructed by using the maximum parsimony method (see Fitch 1977). The trees obtained for the four genes are presented in figures 1 and
Table 1
Influenza A Virus Gene Sequences Used in the Present Study

<table>
<thead>
<tr>
<th>Genes and Subtypes*</th>
<th>Strainsb,c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemagglutinin d</td>
<td></td>
</tr>
<tr>
<td>H1 subtype: SP(48), HAI(228)</td>
<td>WIS/30(S), PR/34, NWS/33, BEL/42, FW/50, LO/57, NJ/76 [1]; WSN/33 [2]; USSR/77 [3]</td>
</tr>
<tr>
<td>H2 subtype: SP(42), HAI(207)</td>
<td>RI/57, TOK/67, NED/68, BER/68, GDR/72(D), ALB/77(D), ALB/77(P) [4]</td>
</tr>
<tr>
<td>H3 subtype: SP(45), HAI(983), HA2(656)</td>
<td>UKR/65(D) [5]; NT/68 [6, 7]; ENG/69, QU/70 [7]; MEM/71 [8]; MEM/72 [9]; HK/71, ENG/72, PC/73, VIC/76, AC/76, TX/77, BA2/79 [10]; VIC/75 [11]; ENG/77 [12]; BA1/79 [13]</td>
</tr>
<tr>
<td>H11 subtype: SP(45), HAI(201)</td>
<td>ENG/56(D), UKR/60(D), AU/75(T), MEM/76(D), NY/78(D) [14]</td>
</tr>
<tr>
<td>Neuraminidase:</td>
<td></td>
</tr>
<tr>
<td>N1 subtype: (168)</td>
<td>WIS/30(S), PR/34, BEL/42, FW/50, LO/57, NJ/76, ON/77(D), USSR/77, MEM/78 [15]; WSN/33 [16]</td>
</tr>
<tr>
<td>N2 subtype: (1404)</td>
<td>RI/57 [17]; NT/68 [18]; UD/72 [19]; VIC/75 [20]; BA1/79 [21]</td>
</tr>
<tr>
<td>Matrix protein 1:</td>
<td>PR/34 [22]; FPV/34(F) [23]; FW/50, LO/57, RI/57, CG/77 [4]; UD/72 [24]; BA1/79 [25]</td>
</tr>
<tr>
<td>Nonstructural protein 1:</td>
<td></td>
</tr>
<tr>
<td>Partial sequences (192)</td>
<td>RI/57, CG/77, AU/78(B) [4]</td>
</tr>
<tr>
<td>Complete sequences (687) f</td>
<td>PR/34 [26]; FPV/34(F) [27]; FM/47, FW/50, USSR/77 [28]; UD/72 [29]; ALB/76 [30]; AL/77 [31]</td>
</tr>
</tbody>
</table>

* Numbers in parentheses are the number of nucleotides.


c The letter in parentheses denotes the organism from which the virus strain was isolated. B = black duck, D = duck, P = pintail (these three belong to Anatinae); T = tern (Sterininae); F = fowl (including chicken and duck); and S = swine (Sus scrofa). Strains without parentheses are from humans.

d SP = signal peptide; HAI = hemagglutinin 1; and HA2 = hemagglutinin 2.

e Ref. 13 does not have SP sequences, and Refs. 7 and 10 have only HAI sequences, except for HA2 sequence of the MEM/72 strain. For NT/68, SP and HA2 are from ref. 6, and HAI is from ref. 7.

f Since the length of open reading frame for three strains (PR/34, FPV/34, and ALB/76) is shorter than that for others, seven unshared codons are not included.

2. When two or more trees with nearly the same total number of substitutions (one or two differences) were obtained for the same set of data, a consensus tree with collapsed node denoted by C was produced (e.g., fig. 1 B). The root of a tree was located by using an outside strain. For hemagglutinin subtypes (fig. 1 A–1 D), Webster et al.'s (1982) dendrogram was used to identify outside strains. In the case of neuraminidase subtypes (fig. 2 A and 2 B), there were no such dendrograms available, but for the N1 subtype the root could be located by using information on the H1 subtype, because most of the strains studied were the same for the H1 and N1 subtypes. For the N2 subtype, the oldest strain (RI/57) was assumed to be the ancestor (root). For matrix protein 1, FPV/34(F) was used as the outside strain, and for nonstructural protein 1, ALB/76(D) was assumed to be the outside strain. Both strains were isolated from
FIG. 1.—Phylogenetic trees for the H1, H2, H3, and H11 subtypes of the hemagglutinin gene reconstructed by the maximum parsimony method. An unresolved node of a consensus tree is denoted by a circle enclosing a C and the root by a circle enclosing an O. For the determination of roots, see text. The letter in parentheses denotes the initial of the organism from which the strain was isolated. Strains without this initial are from humans. For the names of organisms, see table 1. Branch lengths are approximately proportional to the number of nucleotide substitutions. (A) = the H1 subtype (both hemagglutinin 1 [HA1] and signal peptide [SP] used); (B) = the H2 subtype (HA1 and SP used); (C) = the H3 subtype (only HA1 used); and (D) = the H11 subtype (HA1 and SP used).

birds and were quite different from other human strains. Since the number of nucleotide substitutions for each branch cannot be estimated uniquely by the maximum parsimony method, only approximate branch lengths are given in figures 1 and 2.

The phylogenetic trees in figures 1 and 2 are quite different from ordinary trees
because the strains used here were isolated in different years and mutations accumulated rapidly. The differences in evolutionary pattern among the phylogenetic trees in figures 1 and 2 largely depend on whether there was polymorphism in the past. If a gene were always monomorphic, one would expect all strains isolated in different years to be located on a line without branching, as in the case of FW/50 and RI/57 in figure 2(C) (matrix protein 1). On the other hand, if a gene is highly polymorphic and polymorphic alleles stay in the population for a long time, we would expect the type of trees represented by figures 1A and 1D (hemagglutinin).
Figures 1(A), 1(B), 1(C), and (D) show the phylogenetic trees for the \(H_1, H_2, H_3,\) and \(H_{11}\) subtypes of the hemagglutinin gene, respectively. A substantial amount of polymorphism exists in all subtypes, and, in general, mutations accumulate more or less linearly with time. However, figure 1(A) shows one anomaly. That is, strain USSR/77 did not accumulate mutations very much after it separated from FW/50. This anomaly has been previously noted by Nakajima et al. (1978) and will be discussed later. Subtypes \(H_4-H_{13}\) have been found only in nonhuman species, but the other three subtypes infect humans as well as other organisms. Except for NJ/76 of the \(H_1\) subtype, strains of subtypes \(H_1, H_2,\) and \(H_3\) isolated from humans are monophyletic and form one cluster (fig. 1A). NJ/76 is known to be close to swine strains (Blok and Air 1982), though it was isolated from humans, and our result, showing this strain clustered with a swine strain (WIS/30), confirms Blok and Air’s earlier conclusion. The recent \(H_3\) subtype strains in humans (Hong Kong flu) are all derived from NT/68, which in turn shares a common ancestor with UKR/63 isolated from duck. A main trunk is noticed in the phylogenetic tree, suggesting that in each year there was a dominant strain. The topology of the tree of \(H_3\) subtype is similar to that of Both et al. (1983) if we compare the same set of 12 strains. The \(H_{11}\) subtype infects only nonhuman species and seems to be highly polymorphic.

Figures 2(A) and 2(B) present the phylogenetic trees for the \(N_1\) and \(N_2\) subtypes of the neuraminidase gene, respectively. The tree for the \(N_1\) subtype shares seven strains with that for the hemagglutinin \(H_1\) subtype (fig. 1A). This is because the subtypes that were dominant in humans until 1957 are \(H_1\) for the hemagglutinin gene and \(N_1\) for the neuraminidase gene. Consequently, the topologies of these two trees are similar. In the tree of the \(N_2\) subtype (fig. 2B), a trunk is identified, as in the case of the \(H_3\) subtype of the hemagglutinin gene.

Figures 2(C) and 2(D) represent phylogenetic trees for the matrix protein 1 (\(M_X_1\)) and nonstructural protein 1 (\(N_S_1\)) genes, respectively. The tree for \(N_S_1\) is similar to that of Buonagurrio et al. (1984) if we consider the same set of strains. The trees for \(M_X_1\) and \(N_S_1\) share six strains, and the topological relationships of these six strains are identical, if the branch linking strains FW/50 and RI/57 is eliminated. This similarity of topology suggests that these two sequences have evolved together, although they are located on two different RNA segments. Note that the strains used for constructing a tree for the matrix protein 1 and nonstructural protein 1 do not have the same subtypes of the hemagglutinin and neuraminidase genes. For example, in figure 2(C) the subtypes of the hemagglutinin and neuraminidase genes for strains PR/34, RI/57, and BA1/79 are \(H_{11}N_1, H_{2}N_2,\) and \(H_{3}N_2,\) respectively. This is because the genes of influenza A virus are segmented and occasionally reassorted (Webster et al. 1982).

**Pattern of Accumulation of Nucleotide Substitutions**

We studied the pattern of nucleotide substitution using the main evolutionary lines identified by the above phylogenetic analysis. The oldest strains, which were at the root (RI/57 for the \(N_2\) subtype) or near the root (PR/34 for \(H_1, N_1, M_X_1,\) and \(N_S_1;\) RI/57 for \(H_2;\) and NT/68 for \(H_3),\) were treated as the ancestral strains, and strains that diverged earlier than the appearance of the ancestral strains were excluded from the analysis to minimize the effect of polymorphism. Most of the strains excluded were those isolated from nonhuman vertebrates, and the strains used for the analysis...
were all from humans. The H11 subtype was not included in the analysis, since no main evolutionary line was identified. The number of nucleotide substitutions per site between two strains was estimated by means of Jukes and Cantor's (1969) formula.

Figure 3 presents the pattern of accumulation of nucleotide substitutions for the genes or subtypes (H1, N1, MX1, and NS1) in which the ancestral strain was identified as PR/34. The results for the other subtypes (H2, H3, and N2) are given in figures 4 and 5. In the H1 subtype, data for SP and HA1 were combined, since they were similar. There are two sets (partial and complete sequences) of data for NS1, and the results from the comparison of complete sequences (687 bases) are shown, the data for strains RI/57, CG/77, and AU/78 being excluded.

The accumulation of nucleotide substitutions is approximately linear for all genes and subtypes examined. Particularly, in the H3 subtype of the HA1 gene, where the largest number of strains is used, the linearity is quite satisfactory (fig. 5). However, there are exceptional strains that do not follow the pattern of linear accumulation. They are USSR/77 and its close relative MEM/78 (marked by open circles in fig. 3). The H1, N1, and NS1 genes in these strains show a small number of nucleotide substitutions after they branched off about 1950. A slowdown of nucleotide substitution
is also observed for the $MXI$ gene of strain CG/77, though it could arise from stochastic errors. All genes of the USSR/77 strain have been shown to be quite similar to those strains circulating about 1950 (Nakajima et al. 1978), and the phylogenetic trees for $H1$, $N1$, and $NSI$ confirm this finding (see Discussion). Since USSR/77 and MEM/78 are clearly abnormal, they were excluded from the following analysis.

**Rate of Nucleotide Substitution**

We applied two types of regression analyses to estimate the rate of nucleotide substitution. One was the regression through the origin as used by Nei (1983). This method seems to be suitable for the $H2$ and $H3$ subtypes of the hemagglutinin gene, the $N2$ subtype of the neuraminidase gene, and the $MXI$ gene, where the ancestral strains of these genes are at the root or quite close to the root. On the other hand, when the ancestral strain branched off the root and if the branch length is not negligibly small as in the case of the $NSI$ gene from strain PR/34, this method may give an overestimate of the substitution rate. Therefore, the usual regression method, which would alleviate this problem, was also used for the $H1$, $N1$, and $NSI$ genes. The regression coefficients for the former and the latter methods are given by

\[ b_1 = \frac{\sum x_i y_i}{\sum x_i^2} \]
FIG. 5.—Accumulation of nucleotide substitutions in the H3 subtype of the hemagglutinin 1 gene of the influenza A virus genes. The patterns of the accumulation of nucleotide substitutions for the first and second positions (open circles) and for the third position (black dots) are shown. The year of isolation of the ancestral strain is 1968. For each case, the regression lines are superimposed. The regression coefficient is 0.0050/site/year for the first and second position and 0.0104/site/year for the third position.

The regression coefficients thus obtained were used as estimates of the rate of substitution. These estimates were obtained for each of the three nucleotide positions of codons separately. The results obtained arc shown in table 2. In this table, the average rates for the first and second positions are combined, since these two positions show similar rates.

Table 2 shows that the rates of nucleotide substitution obtained from $b_2$ are generally smaller than those obtained from $b_1$, as expected. The rate of nucleotide substitution varies considerably with the gene examined. The highest rate (0.0124) for "all positions" is observed for the signal peptide of the H3 subtype, and the lowest

respectively, where $x_i$ is the time of isolation of the $i$th strain measured from the origin, $y_i$ is the number of nucleotide substitutions between the $i$th strain and the ancestral strain, and $n$ is the number of strains used excluding the ancestral one. (Note that $b_1$ usually has a smaller sampling error than does $b_2$ because the number of degrees of freedom for $b_1$ is larger than that for $b_2$.) The regression coefficients thus obtained were used as estimates of the rate of substitution. These estimates were obtained for each of the three nucleotide positions of codons separately. The results obtained are shown in table 2. In this table, the average rates for the first and second positions are combined, since these two positions show similar rates.

Table 2 shows that the rates of nucleotide substitution obtained from $b_2$ are generally smaller than those obtained from $b_1$, as expected. The rate of nucleotide substitution varies considerably with the gene examined. The highest rate (0.0124) for "all positions" is observed for the signal peptide of the H3 subtype, and the lowest
Table 2
Rates of Nucleotide Substitution in Influenza A Virus Genes

<table>
<thead>
<tr>
<th>GENE</th>
<th>SUBTYPE</th>
<th>N</th>
<th>1st + 2d</th>
<th>3d</th>
<th>All</th>
<th>RATIO (1st + 2d/3d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal peptide:</td>
<td>H1</td>
<td>(4)</td>
<td>0.29 ± 0.11</td>
<td>1.14 ± 0.16</td>
<td>0.55 ± 0.08</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(−0.01 ± 0.15)</td>
<td>(1.16 ± 0.23)</td>
<td>(0.47 ± 0.11)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>H2</td>
<td>(4)</td>
<td>0.35 ± 0.20</td>
<td>1.48 ± 0.05</td>
<td>0.71 ± 0.14</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>H3</td>
<td>(4)</td>
<td>1.39 ± 0.21</td>
<td>0.95 ± 0.27</td>
<td>1.24 ± 0.23</td>
<td>1.46</td>
</tr>
<tr>
<td>Mean(^b)</td>
<td></td>
<td></td>
<td>0.68</td>
<td>1.18</td>
<td>0.83</td>
<td>0.58</td>
</tr>
<tr>
<td>Hemagglutinin 1:</td>
<td>H1</td>
<td>(4)</td>
<td>0.29 ± 0.06</td>
<td>0.91 ± 0.13</td>
<td>0.48 ± 0.08</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.10 ± 0.08)</td>
<td>(0.54 ± 0.18)</td>
<td>(0.23 ± 0.11)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>H2</td>
<td>(4)</td>
<td>0.12 ± 0.02</td>
<td>1.08 ± 0.23</td>
<td>0.42 ± 0.07</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>H3</td>
<td>(15)</td>
<td>0.50 ± 0.02</td>
<td>1.04 ± 0.03</td>
<td>0.68 ± 0.01</td>
<td>0.48</td>
</tr>
<tr>
<td>Mean(^b)</td>
<td></td>
<td></td>
<td>0.41</td>
<td>1.02</td>
<td>0.61</td>
<td>0.40</td>
</tr>
<tr>
<td>Hemagglutinin 2:</td>
<td>H3</td>
<td>(6)</td>
<td>0.10 ± 0.02</td>
<td>0.70 ± 0.07</td>
<td>0.30 ± 0.03</td>
<td>0.14</td>
</tr>
<tr>
<td>Neuraminidase:</td>
<td>N1</td>
<td>(4)</td>
<td>0.38 ± 0.01</td>
<td>0.61 ± 0.09</td>
<td>0.45 ± 0.04</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.39 ± 0.02)</td>
<td>(0.80 ± 0.12)</td>
<td>(0.52 ± 0.05)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N2</td>
<td>(5)</td>
<td>0.29 ± 0.02</td>
<td>0.63 ± 0.03</td>
<td>0.40 ± 0.02</td>
<td>0.46</td>
</tr>
<tr>
<td>Mean(^b)</td>
<td></td>
<td></td>
<td>0.30</td>
<td>0.63</td>
<td>0.41</td>
<td>0.48</td>
</tr>
<tr>
<td>Nonstructural protein 1</td>
<td></td>
<td>(5)</td>
<td>0.17 ± 0.02</td>
<td>0.44 ± 0.04</td>
<td>0.26 ± 0.03</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.10 ± 0.02)</td>
<td>(0.31 ± 0.05)</td>
<td>(0.17 ± 0.03)</td>
<td></td>
</tr>
<tr>
<td>Matrix protein 1:</td>
<td></td>
<td>(7)</td>
<td>0.05 ± 0.01</td>
<td>0.37 ± 0.04</td>
<td>0.15 ± 0.01</td>
<td>0.14</td>
</tr>
<tr>
<td>Grand mean(^b)</td>
<td></td>
<td></td>
<td>0.31</td>
<td>0.71</td>
<td>0.41</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Values of \(\lambda\) and its SE are multiplied by 100. Figures in parentheses are \(b_2\) and its SE. 1st, 2d, and 3d denote the first, second, and third nucleotide positions of codons, respectively.

Mean of \(b_2\) weighted by the number of nucleotides.

Rate (0.0015) is observed in the gene for matrix protein 1, the former being about eight times higher than the latter when all nucleotide positions were compared. The rate for the signal peptide of the \(H3\) subtype is exceptionally high, and the other subtypes for the signal peptide show lower rates. This difference results mainly from the high rate in the first and second positions for the \(H3\) subtype. Similar difference among subtypes is observed in the hemagglutinin 1 gene. However, even the average rate (0.0083/site/year) for three subtypes of the signal peptide is more than five times higher than that of the matrix protein 1 gene. The average rate for all genes examined is 0.0041/site/year.

Martinez et al. (1983) estimated the rate of nucleotide substitution for the \(N2\) subtype of the neuraminidase gene to be 0.0033–0.0070. Our estimate for this subtype (0.0040) is closer to their minimum estimate. Krystal et al.'s (1983) estimates for nonstructural proteins and for the \(H3\) subtype of hemagglutinins 1 and 2 are 0.0022–0.0034 and 0.0045–0.0065, respectively. Our estimates for the \(NSI\) gene (0.0026 by \(b_2\)) and for the hemagglutinin gene (the average of estimates for \(HA1\) and \(HA2\) of the \(H3\) subtype weighted by the number of nucleotides = 0.0053) are again closer to the lower ends of the range of their estimates. This indicates that the estimate obtained
from a simple comparison of the two strains tends to give an overestimate. Nei's (1983) estimate for $MXI$ (0.0013) and $NSI$ (0.0027) are both close to the estimates obtained in the present study. In the case of $NSI$, Nei (1983) used Air and Hall's (1981) shorter (192 nucleotides) sequence data for four strains, whereas we used longer (687 nucleotides) sequence data (see table 1) for five strains. The agreement between the two estimates suggests that the rate of nucleotide substitution for $NSI$ is rather uniform for all gene regions.

It is seen from table 2 that the average rate of nucleotide substitution for the first and second nucleotide positions of codons is considerably lower than that for the third position except for the $H3$ signal peptide. Indeed, the ratio of the former to the latter is 0.62 or less. This is because certain amino acids in proteins are highly conserved in the evolutionary process. This is known to be the case even with this rapidly evolving virus (Nei 1983).

The above estimates of the rate of nucleotide substitution were obtained using human strains only. It is difficult to obtain reliable estimates of the substitution rate for strains derived from nonhuman vertebrates, but there are some indications that the rate for nonhuman strains is lower than that for human strains. For example, strain NJ/76 and a swine strain WIS/30 share a common ancestor in both the $HI$ subtype of $HAI$ (fig. 1A) and the $NI$ subtype of neuraminidase (fig. 2A). (NJ/76 is known to be essentially a swine strain, though it was isolated from humans; Blok and Air 1982.) Comparison of these two strains is expected to give an overestimate of the rate, but the estimates obtained (0.0024 and 0.0031 for $HI$ and $NI$, respectively) are lower than those obtained from human strains. All five strains of the $H11$ subtype of hemagglutinin were derived from birds. Because there was no main evolutionary line for this gene (fig. 1D), these strains were not used for estimating substitution rate. However, if we use MEM/76 as a reference strain, the rate can be estimated by the regression analysis ($b_2$) mentioned earlier. This gives an estimate of 0.0019 (the signal peptide and $HAI$ genes combined), which is approximately one-third of the average rate for three $HAI$ genes. Comparison of two bird strains, FPV/34(F) and AU/78(B), in the $NSI$ gene (fig. 2D) also gives a rate (0.0008) that is approximately one-third to one-half of the rate (0.0026 by $b_1$ or 0.0017 by $b_2$) for human strains (table 2). It is not clear why the rate of nucleotide substitution is lower in nonhuman viral strains than in human strains, if this difference is real.

Discussion

Rate of Nucleotide Substitution in RNA Genomes

We have seen that the rate of nucleotide substitution is of the order of $10^{-3}$/site/year for most influenza A virus genes studied. Using the oligonucleotide mapping technique, Takeda et al. (1984) estimated the substitution rate for the enterovirus type 70 (single-stranded RNA virus) to be $4 \times 10^{-3}$/site/year. Gojobori and Yokoyama (1985) estimated that the substitution rates for the $gag$ gene (gene for internal proteins) and for the $v-mos$ gene (oncogene) in a retrovirus (single-stranded RNA virus) are $6.3 \times 10^{-4}$ and $1.31 \times 10^{-3}$, respectively. These high rates of nucleotide substitution in RNA viruses are believed to arise from the absence of proofreading exonucleases for correcting replication error (Holland et al. 1982). In DNA viruses such as the papovavirus, the rate of nucleotide substitution is much lower and of the order of $10^{-9}$ (Soeda and Maruyama 1982). The same order of substitution rate $[(2-4) \times 10^{-8}]$
has been observed for many mammalian genes (Li et al. 1985). Since DNA viruses use the replication system of the host, the mutation rate is expected to be similar to that of the host.

We have seen that the rate of nucleotide substitution at the first and second positions of codons is much lower than that at the third position except in the gene for the H3 signal peptide of hemagglutinin. In this exceptional gene, all three nucleotide positions have essentially the same substitution rate, and the rate is \( \sim 0.01 \text{/site/year} \). These values suggest that for some reason there is little purifying selection operating for this gene and that nucleotide substitution occurs at the same rate as does the mutation rate. They also suggest that the relatively high substitution rates for HAI and neuraminidase are not caused by positive selection pressure owing to the host immune system against this virus but are simply the result of a high mutation rate. The lower rate for HA2 than for HAI apparently occurs because hemagglutinin 2 constitutes the supporting leg for hemagglutinin 1 (Wilson et al. 1981) and thus is subject to stronger purifying selection than hemagglutinin 1. The lowest rate observed for the matrix protein 1 gene is also understandable from the function of this protein, since this protein underlies the lipid bilayer of the viral core and is important for viral construction and budding (Lamb 1983). Generally speaking, there seems to be a good correlation between the rate of nucleotide substitution and the level of functional constraint in the influenza A virus genes, as is true in the case of DNA genomes (Kimura 1983).

Deceleration of Substitution Rate

Since Nakajima et al. (1978) reported the striking similarity between USSR/77 and the strains that were circulating during the 1950s, the problem of deceleration of nucleotide substitution has attracted attention from many authors. Krystal et al. (1983) sequenced NS1 genes from two strains isolated about 1950 and compared these sequences with that of USSR/77. They showed that only five nucleotide substitutions occurred during more than 20 years. Our lower estimate (via \( b_2 \)) of the substitution rate for the NS1 gene is 0.0017 (table 2). Therefore, if nucleotide substitution occurred at this rate, there should have been approximately 30 substitutions from 1950 to 1977 in the sequence of 687 nucleotides. This indicates how slow the nucleotide substitution was in the USSR/77 lineage. The same slow rate of substitution is also observed for the MEM/78 strain of the N1 subtype of neuraminidase (see fig. 3). The reason for this unusual deceleration is not known at present. The deceleration of the rate does not seem to be an inherent property of these strains, since after their reappearance they started to evolve quickly (Young et al. 1979). Palese and Young (1983) speculated that USSR/77 might have originated from a laboratory stock that had been kept frozen for a long time. Another possibility is that about 1950 the ancestor of this strain infected some nonhuman organism, one in which the replication rate of this virus was reduced, and that it later reinfected humans.

Age of Sequence Polymorphism

The phylogenetic trees in figures 1 and 2 indicate that human and nonhuman strains can coexist for a long time showing sequence polymorphism. For example, the polymorphism of strains (or lineages) NJ/76 (a "swine" strain) and LOY/57 in H1 (fig. 1A) apparently existed for more than 27 years (i.e., from 1930 to 1957). It is
possible to estimate the time of divergence between these two strains under the assumption of a constant rate of evolution. Let \( t_1 \) and \( t_2 \) be the years of isolation of strains 1 and 2 respectively, with \( t_1 > t_2 \). Then, the expected number of nucleotide substitutions between the two strains (\( d \)) may be expressed as 
\[
    d = \frac{2T + t_1 - t_2}{2\lambda},
\]
where \( T \) is the time of duration of polymorphism, i.e., the time between \( t_2 \) and the year of divergence (the root denoted by \( O \)), and \( \lambda \) is the rate of nucleotide substitution per site per year. The \( d \) value may be computed by Jukes and Cantor's (1969) formula. Therefore, \( T \) can be estimated by the equation
\[
    T = \frac{d}{2\lambda} - \frac{t_1 - t_2}{2}.
\]

In the case of the HI subtype (fig. 1A), we do not know the \( \lambda \) value for nonhuman strains, but if we use \( \lambda = 0.005 \) (the average for SP and HAI) from table 2, we obtain \( T = 16 \) years. This value is certainly an underestimate and apparently caused by the fact that \( \lambda \) is smaller in nonhuman vertebrates than in humans. In the case of non-structural protein 1 in figure 2(D), a more reasonable result is obtained even if we use the \( \lambda \) value from human viral strains. In this case the polymorphism of AL/77 (human strain) and ALB/76 (duck strains) existed for at least 42 years (i.e., from 1934 to 1976), and the \( d \) value between the two strains is 0.383. If we use \( \lambda = 0.0026 \), we obtain \( T = 72 \). This suggests that the two strains have coexisted since 1904 or earlier.

Polymorphism of Hemagglutinin Genes

As mentioned earlier, 13 different subtypes have been identified in the hemagglutinin gene, and the extent of nucleotide divergence is very large. Nevertheless, they are polymorphic "alleles" in classical genetics and have the same biological function. It is known that they are usually host specific, and three of them (HI, H2, and H3) are carried by strains infecting humans. However, host specificity is not absolute, and switching of hosts occurs occasionally (e.g., Webster et al. 1982). The high degree of genetic diversity among the subtypes seems to result largely from the high rate of mutation in this virus in association with host specificity (Nei 1983). It is therefore interesting to know the evolutionary history of these subtypes. For this purpose we constructed a dendrogram for 13 subtypes of this gene by using the unweighted pair-grouping method (UPGMA). Previously, Hinshaw et al. (1982) and Webster et al. (1982) constructed similar dendrograms, but they used various strains that were isolated in different years. In our study we used only one strain for each subtype and chose the one that was isolated in the same or nearly the same year. In this study we used amino acid sequences rather than nucleotide sequences because the former give more reliable results when genetic divergence is large. We estimated the number of amino acid replacements per site for each pair of subtypes by the equation
\[
    d = -\log_2(1 - p),
\]
where \( p \) is the proportion of different amino acids. In this computation we used a region of 78 amino acids (the shared segment of hemagglutinin 1 excluding deletions/insertions; see Air 1981 and Hinshaw et al. 1982) but excluded 10 amino acids that are apparently invariable. Thus, a total of 68 amino acids were used. The proportion of different amino acids (\( p \)) and the number of amino acid replacements per site (\( d \)) for all pairs of subtypes are given in table 3, and the dendrogram obtained by UPGMA from this \( d \) matrix is presented in figure 6. This dendrogram need not represent the
Table 3
Amino Acid Differences among 13 Hemagglutinin Subtypes

<table>
<thead>
<tr>
<th></th>
<th>H1</th>
<th>H2</th>
<th>H5</th>
<th>H11</th>
<th>H6</th>
<th>H13</th>
<th>H8</th>
<th>H9</th>
<th>H12</th>
<th>H7</th>
<th>H10</th>
<th>H4</th>
<th>H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>0.28</td>
<td>0.34</td>
<td>0.47</td>
<td>0.40</td>
<td>0.57</td>
<td>0.62</td>
<td>0.54</td>
<td>0.60</td>
</tr>
<tr>
<td>H2</td>
<td>0.33</td>
<td>0.22</td>
<td>0.44</td>
<td>0.44</td>
<td>0.54</td>
<td>0.60</td>
<td>0.53</td>
<td>0.59</td>
<td>0.72</td>
<td>0.75</td>
<td>0.82</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>H5</td>
<td>0.41</td>
<td>0.25</td>
<td>0.46</td>
<td>0.57</td>
<td>0.62</td>
<td>0.56</td>
<td>0.57</td>
<td>0.74</td>
<td>0.74</td>
<td>0.81</td>
<td>0.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H11</td>
<td>0.64</td>
<td>0.58</td>
<td>0.58</td>
<td>0.44</td>
<td>0.43</td>
<td>0.56</td>
<td>0.53</td>
<td>0.56</td>
<td>0.78</td>
<td>0.82</td>
<td>0.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H6</td>
<td>0.51</td>
<td>0.58</td>
<td>0.61</td>
<td>0.58</td>
<td>0.65</td>
<td>0.57</td>
<td>0.56</td>
<td>0.54</td>
<td>0.75</td>
<td>0.76</td>
<td>0.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H13</td>
<td>0.85</td>
<td>0.79</td>
<td>0.85</td>
<td>0.56</td>
<td>1.04</td>
<td>0.54</td>
<td>0.54</td>
<td>0.60</td>
<td>0.78</td>
<td>0.82</td>
<td>0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H8</td>
<td>0.96</td>
<td>0.92</td>
<td>0.96</td>
<td>0.82</td>
<td>0.85</td>
<td>0.79</td>
<td>0.35</td>
<td>0.40</td>
<td>0.78</td>
<td>0.76</td>
<td>0.81</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>H9</td>
<td>0.79</td>
<td>0.75</td>
<td>0.82</td>
<td>0.75</td>
<td>0.82</td>
<td>0.79</td>
<td>0.44</td>
<td>0.35</td>
<td>0.78</td>
<td>0.79</td>
<td>0.75</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>H12</td>
<td>0.92</td>
<td>0.89</td>
<td>0.85</td>
<td>0.82</td>
<td>0.79</td>
<td>0.92</td>
<td>0.51</td>
<td>0.44</td>
<td>0.78</td>
<td>0.78</td>
<td>0.81</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>H7</td>
<td>1.45</td>
<td>1.28</td>
<td>1.33</td>
<td>1.51</td>
<td>1.39</td>
<td>1.51</td>
<td>1.51</td>
<td>1.51</td>
<td>0.50</td>
<td>0.68</td>
<td>0.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H10</td>
<td>1.33</td>
<td>1.39</td>
<td>1.33</td>
<td>1.74</td>
<td>1.45</td>
<td>1.45</td>
<td>1.45</td>
<td>1.58</td>
<td>0.69</td>
<td>0.62</td>
<td>0.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H4</td>
<td>1.74</td>
<td>1.74</td>
<td>1.66</td>
<td>1.74</td>
<td>1.45</td>
<td>1.45</td>
<td>1.45</td>
<td>1.66</td>
<td>1.13</td>
<td>0.96</td>
<td>0.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3</td>
<td>1.66</td>
<td>1.51</td>
<td>1.51</td>
<td>1.66</td>
<td>1.45</td>
<td>2.27</td>
<td>2.14</td>
<td>1.82</td>
<td>1.66</td>
<td>0.96</td>
<td>0.89</td>
<td>0.89</td>
<td></td>
</tr>
</tbody>
</table>

NOTE.—Figures above the diagonal are the observed proportions of amino acid differences (p) in a sequence of 68 amino acids compared. Figures below the diagonal are the estimated number (d) of amino acid replacements per site. Subtypes are arranged in the order of appearance in the dendrogram of fig. 6.

true evolutionary tree, since the rate of amino acid replacements might vary with host and host switching occasionally occurs. Nevertheless, it gives a rough idea of the evolutionary divergence of the 13 subtypes of the hemagglutinin gene. Essentially the same tree topology was obtained by using Farris's (1972) distance Wagner method (tree not shown).

The topology of the dendrogram obtained is quite similar to those of Hinshaw et al. (1982) and Webster et al. (1982), though these authors used several strains for

![Phylogenetic tree of 13 hemagglutinin subtypes of the influenza A virus made by UPGMA.](image)

**Fig. 6.**—Phylogenetic tree of 13 hemagglutinin subtypes of the influenza A virus made by UPGMA. Sixty-eight amino acid sequences are used. H1 = New Jersey/11/76 (Air et al. 1981); H2 = duck/Alberta/77/77 (Air and Hall 1981); H3 = Texas/1/77 (Webster et al. 1982); H4 = duck/Alberta/28/76 (Air 1981); H5 = shearwater/Australia/28/76 (Air 1981); H6 = shearwater/Australia/72 (Air 1981); H7 = turkey/Oregon/71 (Air 1981); H8 = duck/Alberta/283/77 (G. M. Air, unpublished data); H9 = turkey/Wisconsin/1/66 (Air 1981); H10 = quail/Italy/1117/65 (G. M. Air, unpublished data); H11 = duck/Memphis/546/76 (Air 1981); H12 = duck/Alberta/60/76 (Air 1981); and H13 = gull/Md/704/77 (Hinshaw et al. 1982).
most subtypes. This is apparently because the intrasubtype divergence is generally much smaller than the intersubtype divergence (Air 1981). However, the branch lengths of our dendrogram are considerably longer than those of the dendrograms of these authors. Air (1981) obtained a similar dendrogram for 12 subtypes of the hemagglutinin gene.

How long did it take for these subtypes to diverge from each other? This is a difficult question to answer, since the rate of amino acid replacement seems to vary with host and the extent of amino acid difference is so high. It is possible that some subtypes have existed for a long time and that the remaining similarity between them results from functional constraint rather than from lack of time to diverge. Nevertheless, it is interesting to know how long it would take for the observed level of divergence to occur under the assumption of a constant rate of amino acid replacement. For this purpose we estimated the rate of amino acid replacement for the H1, H2, H3, and H11 subtypes using the same method ($b_2$) as that used to estimate nucleotide substitution. The estimates obtained were 0.0010, 0.0005, 0.0102, and 0.0009/amino acid site/year for subtypes H1, H2, H3, and H11, respectively, the average being 0.0035. If we accept this replacement rate, the divergence between the two largest clusters of hemagglutinins in figure 6 corresponds to 200–300 years. One may compare this result with the age of polymorphic alleles at the alcohol dehydrogenase locus in *Drosophila melanogaster*. In this case the average nucleotide difference between two randomly chosen alleles (DNA sequences) is only 0.007, but the coalescence time of polymorphic alleles (time of the earliest divergence) has been estimated to be $\sim$1 Myr (Stephens and Nei 1986). This indicates how rapidly gene divergence has occurred in the influenza A virus.

Note added in proof.—After the submission of this paper, Hayashida et al. (Mol. Biol. Evol. 2:289–303, 1985) reported an analysis of influenza A virus genes. Their general conclusions (extremely high mutation rate and constant rate of evolution) are virtually the same as ours as well as Nei’s (1983). There is, however, one notable difference. Hayashida et al. located the Sage (A/duck/Ontario/77) strain as a direct descendant of a 1950 strain (FW) in their figure 4. In our analysis, however, FW [=FW/50] is not the ancestor of the Sage [=ON/77(D)] strain but both are derived from a common ancestor (fig. 2A). Therefore, Hayashida et al.’s conclusion (9 years of frozen period) for the Sage strain is not supported by our analysis.

Acknowledgments

We thank Dr. G. M. Air for sending us two unpublished sequences of *HAI* subtypes. We also thank Drs. A. K. Roychoudhury, J. C. Stephens, D. Graur, T. Gojobori, P. Pamilo, and W. M. Fitch for their comments. This study was supported by research grants from the National Institutes of Health and the National Science Foundation to M. Nei.

LITERATURE CITED


Walter M. Fitch, reviewing editor

Received April 18, 1985; revision received August 1, 1985.
Molecular Evolution of Bacteriophages: Evidence of Selection against the Recognition Sites of Host Restriction Enzymes

Paul M. Sharp
Department of Genetics, Trinity College, Dublin, Ireland

Restriction enzymes produced by bacteria serve as a defense against invading bacteriophages, and so phages without other protection would be expected to undergo selection to eliminate recognition sites for these enzymes from their genomes. The observed frequencies of all restriction sites in the genomes of all completely sequenced DNA phages (T7, lambda, φX174, G4, M13, f1, fd, and IKe) have been compared to expected frequencies derived from trinucleotide frequencies. Attention was focused on 6-base palindromes since they comprise the typical recognition sites for type II restriction enzymes. All of these coliphages, with the exception of lambda and G4, exhibit significant avoidance of the particular sequences that are enterobacterial restriction sites. As expected, the sequenced fraction of the genome of φ29, a Bacillus subtilis phage, lacks Bacillus restriction sites. By contrast, the RNA phage MS2, several viruses that infect eukaryotes (EBV, adenovirus, papilloma, and SV40), and three mitochondrial genomes (human, mouse, and cow) were found not to lack restriction sites. Because the particular palindromes avoided correspond closely with the recognition sites for host enzymes and because other viruses and small genomes do not show this avoidance, it is concluded that the effect indeed results from natural selection.

Introduction

Bacterial restriction endonucleases, cleaving foreign DNA at or near specific short recognition sequences, function as a defense against infection by parasites (bacteriophages). As with any host-parasite relationship, natural selection should favor the evolution of mechanisms by which phages avoid restriction. Indeed, several such mechanisms exist (Kruger and Bickle 1983), including DNA modification, production of proteins that inhibit the action of certain restriction systems, and use of unusual bases in phage DNA. It has also been suggested that phage genomes contain fewer restriction sites than expected—natural selection should eliminate phages containing these sequences, leading to avoidance of these sites appearing to be an evolutionary strategy. Various surveys of the number of fragments produced by cutting phage DNA with different enzymes have suggested that this may be true in, for example, T7 (Rosenberg et al. 1979), SP01 (Pero et al. 1979), φ29 (Ito and Roberts 1979), and φ1 (Kawamura et al. 1981). Although these results are indicative of an effect, the surveys are inconclusive in two ways. First, not all known host enzymes have been tried. Second, the expected number of sites has been calculated simply from the base com-

1. Key words: bacteriophages, restriction endonucleases, restriction sites, molecular evolution, trinucleotides, palindromes.

Address for correspondence and reprints: Dr. Paul M. Sharp, Department of Genetics, Trinity College, Dublin 2, Ireland.

© 1986 by The University of Chicago. All rights reserved.
0737-4038/86/0301-3106$02.00
position of the genome, ignoring any higher-order interactions between bases. Surveys of the frequencies of the simplest such higher structures, dinucleotides, have shown that they are far from random (Nussinov 1984). It is possible to take account of dinucleotide frequencies—and to survey for all known recognition sites—if sequence data are available.

Adams and Rothman (1982) examined the DNA sequences of the complete genomes of three of the phages (φX174, G4, and fd—all of which are single-stranded DNA coliphages) for which such data were first available. They compiled observed frequencies of sites for nine enzymes derived from species of Enterobacteriaceae, calculated expected frequencies by taking account of not only base composition but also 2nd-order Markov chain effects (trinucleotide frequencies), and concluded that “there is no evidence that reduction in the number of restriction sites has been a significant adaptive strategy in response to host-controlled restriction systems.” On the other hand, a brief examination of the frequencies of 17 enterobacterial restriction sites in the complete genome of the double-stranded DNA bacteriophage T7 suggests that these sites are rarer than expected (Sharp et al. 1985), although expectations were derived simply from base composition. Thus the evidence so far is equivocal.

Several more complete genomes from bacteriophages have been determined, as well as many viruses that infect eukaryotes and would not be expected to avoid restriction sites per se (and so serve as a comparison). Also, many more restriction endonucleases have been described—with new specificities—and isolated from a broader range of bacteria. Here these data have been combined in a more complete analysis. The results indicate that most bacteriophages (including two of the three examined by Adams and Rothman) exhibit a marked avoidance of certain short palindromes. This is attributed to natural selection against the occurrence of restriction sites, since (1) the avoided sequences correspond closely with those recognized by restriction enzymes from the host and (2) genomes other than bacteriophages do not show this avoidance. There are two interesting exceptions, namely lambda and G4.

Material and Methods

The observed frequency of occurrence of all restriction-enzyme recognition sites has been compared to that expected, in all bacteriophages for which the complete genome DNA sequence has been determined.

The available sequences are T7, lambda, φX174 and its relative G4, and four related filamentous phages (M13, fd, f1, and IKe). These are all coliphages, and so a fragment comprising approximately one-third of the genome of φ29, a phage isolated from Bacillus subtilis, was also examined. For comparison, the RNA genome of MS2 (a coliphage), the DNA genomes of four eukaryotic viruses of various lengths (Epstein Barr virus, adenovirus 2, papilloma virus, and SV40), and three mammalian mitochondrial DNA molecules were examined.

Restriction enzymes have been classified into three types (Yuan 1981). Comparatively few type I and type III enzymes have been categorized (see table 1), but a large number of type II enzymes (familiar through their widespread use in molecular genetic manipulation and typified by EcoRI) that recognize a large number of different specific sites (Roberts 1985) have been isolated. These sites are typically short palindromes 4–6 bases in length. Type II restriction enzymes isolated from members of the Enterobacteriaceae, to which the coliphage may have been exposed, are listed in table 2, together with their recognition sites. Most of these enzymes recognize 6-base
Table 1
Frequencies of Type I and Type III Restriction Sites in Coliphage

<table>
<thead>
<tr>
<th>ENZYME</th>
<th>T7</th>
<th>φX174</th>
<th>G4</th>
<th>M13</th>
<th>fd</th>
<th>fl</th>
<th>IKe</th>
<th>lambda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eco A</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GAG(N)GTCA</td>
<td>(7.6)</td>
<td>(0.8)</td>
<td>(1.2)</td>
<td>(0.7)</td>
<td>(0.7)</td>
<td>(0.7)</td>
<td>(0.5)</td>
<td>(4.9)</td>
</tr>
<tr>
<td>Eco B</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>TGA(N)GTGC</td>
<td>(8.3)</td>
<td>(1.9)</td>
<td>(1.4)</td>
<td>(1.2)</td>
<td>(1.3)</td>
<td>(1.3)</td>
<td>(1.5)</td>
<td>(10.6)</td>
</tr>
<tr>
<td>Eco D</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>TTA(N)GTCC</td>
<td>(11.8)</td>
<td>(1.1)</td>
<td>(1.2)</td>
<td>(1.3)</td>
<td>(1.4)</td>
<td>(1.4)</td>
<td>(1.2)</td>
<td>(6.3)</td>
</tr>
<tr>
<td>Eco DXI</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>ATCA(N)ATTTC</td>
<td>(1.2)</td>
<td>(0.2)</td>
<td>(0.2)</td>
<td>(0.5)</td>
<td>(0.4)</td>
<td>(0.4)</td>
<td>(0.5)</td>
<td>(2.4)</td>
</tr>
<tr>
<td>Eco K</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>AAC(N)GTGC</td>
<td>(4.7)</td>
<td>(0.4)</td>
<td>(0.8)</td>
<td>(0.3)</td>
<td>(0.4)</td>
<td>(0.3)</td>
<td>(0.7)</td>
<td>(6.0)</td>
</tr>
<tr>
<td>Type III:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eco P1</td>
<td>126</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>49</td>
</tr>
<tr>
<td>AGACC</td>
<td>(98.7)</td>
<td>(10.8)</td>
<td>(10.4)</td>
<td>(9.7)</td>
<td>(10.4)</td>
<td>(9.7)</td>
<td>(10.7)</td>
<td>(70.6)</td>
</tr>
<tr>
<td>Eco P15</td>
<td>36</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>72</td>
</tr>
<tr>
<td>CAGCAG</td>
<td>(27.5)</td>
<td>(5.5)</td>
<td>(4.6)</td>
<td>(3.0)</td>
<td>(3.3)</td>
<td>(3.2)</td>
<td>(4.7)</td>
<td>(72.3)</td>
</tr>
</tbody>
</table>

NOTE.—Numbers in parentheses are expected frequencies.

palindromes, so that 21 of the 64 possible 6-base palindromes are known to be sites. Type II enzymes isolated from Bacillus species are also shown in table 2. These sites are mainly 4-base and 6-base palindromes. Of the 64 6-base palindromes, 14 are recognized by enzymes isolated from Bacillus subtilis and a further 10 by enzymes from other Bacillus species.

DNA Sequence Data

The T7 sequence was obtained on tape from J. Dunn (see Dunn and Studier 1983) and later modified (see Moffat et al. 1984). The IKe sequence was taken from Peeters et al. (1985). All other sequences were taken from Release 4 of the EMBL Nucleotide Sequence Data Library.

Calculation of Expected Frequencies of Restriction Sites

The product of the genomic frequencies of constituent bases is not an appropriate estimate of the expected frequency of a restriction site. Frequencies of the 16 dinucleotides vary widely between different phylogenetic groups (Nussinov 1984). This variation is not explicable in terms of base composition (Smith et al. 1983). It is of particular relevance here that nonrandom doublet frequencies have been recorded for T7 (Sharp et al. 1985) and other phages (Grantham et al. 1985). McClelland (1985), also examining coliphage genomes, found that the best predictors of tetranucleotide frequencies are trinucleotide frequencies. In the present report expected frequencies of restriction sites have been derived from observed trinucleotide frequencies; for example, Exp GAATTC = p(GAA) × p(T|AA) × p(T|AT) × p(C|TC), where p(GAA) is the probability (i.e., observed frequency) of finding the GAA trinucleotide in the genome examined and p(T|AA) is the probability of finding 'T' after the dinucleotide
Table 2
Recognition Sites for Type II Enzymes Derived from Enterobacteriaceae and Bacillus species

<table>
<thead>
<tr>
<th>Source and Site</th>
<th>Enzyme</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Enterobacteriaceae:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GGCC</td>
<td><em>PstII</em></td>
<td>10</td>
</tr>
<tr>
<td>CCWGG</td>
<td><em>EcoR11</em></td>
<td>1 (4, 7)</td>
</tr>
<tr>
<td>GGNCC</td>
<td><em>Eco3911</em></td>
<td>1 (4)</td>
</tr>
<tr>
<td>AAGCTT</td>
<td><em>EcoVIII</em></td>
<td>1</td>
</tr>
<tr>
<td>RCGGGY</td>
<td><em>Cfr101</em></td>
<td>4</td>
</tr>
<tr>
<td>AGCGCT</td>
<td><em>Eco47111</em></td>
<td>1</td>
</tr>
<tr>
<td>CAGCTG</td>
<td><em>Cfr61</em></td>
<td>4 (9)</td>
</tr>
<tr>
<td>CCWWGG</td>
<td><em>EcoT14</em></td>
<td>1 (3)</td>
</tr>
<tr>
<td>CCCGGG</td>
<td><em>Sma1</em></td>
<td>8 (4)</td>
</tr>
<tr>
<td>CGGCGG</td>
<td><em>Ecl1</em></td>
<td>7</td>
</tr>
<tr>
<td>CGATCG</td>
<td><em>Pvu1</em></td>
<td>9</td>
</tr>
<tr>
<td>YGGCCR</td>
<td><em>Eae1</em></td>
<td>6 (4, 10)</td>
</tr>
<tr>
<td>CTGCAG</td>
<td><em>Eco3611</em></td>
<td>1 (6)</td>
</tr>
<tr>
<td>GAATTCC</td>
<td><em>EcoRI</em></td>
<td>1</td>
</tr>
<tr>
<td>GAGCTC</td>
<td><em>EcoCR1</em></td>
<td>1 (9)</td>
</tr>
<tr>
<td>GATATC</td>
<td><em>EcoRV</em></td>
<td>1</td>
</tr>
<tr>
<td>GGYRCC</td>
<td><em>Eco501</em></td>
<td>1</td>
</tr>
<tr>
<td>GRGNCYC</td>
<td><em>Eco241</em></td>
<td>1 (10)</td>
</tr>
<tr>
<td>GGTACC</td>
<td><em>Kpn1</em></td>
<td>5 (10)</td>
</tr>
<tr>
<td>TCGCGA</td>
<td><em>Sbo13</em></td>
<td>2</td>
</tr>
<tr>
<td><strong>Bacillus species:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCGG</td>
<td><em>Bsu11921</em></td>
<td>1</td>
</tr>
<tr>
<td>CGCG</td>
<td><em>Bsu6633</em></td>
<td>1 (4)</td>
</tr>
<tr>
<td>GGCC</td>
<td><em>BsuR1</em></td>
<td>1 (3, 8)</td>
</tr>
<tr>
<td>GATC</td>
<td><em>BssGII</em></td>
<td>3 (4)</td>
</tr>
<tr>
<td>CCWGG</td>
<td><em>BstGII</em></td>
<td>3</td>
</tr>
<tr>
<td>GCAGC</td>
<td><em>Bbv1</em></td>
<td>6</td>
</tr>
<tr>
<td>GGWCC</td>
<td><em>BamNX</em></td>
<td>2</td>
</tr>
<tr>
<td>ATCGAT</td>
<td><em>BanIII</em></td>
<td>5</td>
</tr>
<tr>
<td>CTGCAG</td>
<td><em>BsuM</em></td>
<td>1 (3)</td>
</tr>
<tr>
<td>CTGCAG</td>
<td><em>Bsu1247</em></td>
<td>1 (4)</td>
</tr>
<tr>
<td>GRGCYCY</td>
<td><em>BanII</em></td>
<td>5 (8, 10)</td>
</tr>
<tr>
<td>GGYRCC</td>
<td><em>BanI</em></td>
<td>5</td>
</tr>
<tr>
<td>GTGCAC</td>
<td><em>Bsp1286</em></td>
<td>8</td>
</tr>
<tr>
<td>GTTACC</td>
<td><em>Bsel</em></td>
<td>3</td>
</tr>
<tr>
<td>GAATGC</td>
<td><em>Rsm1</em></td>
<td>3</td>
</tr>
</tbody>
</table>


NOTE.—Not all enzymes are given, but all recognition sites are covered. Thus only one of a set of isoschizomers appears, and where 6-base sites are cut by an enzyme recognizing the 4-base core site, enzymes recognizing only the former are omitted.


AA or p(AAT)/p(AA). For illustration, the effect of using such predictors for 6-base palindromes in T7 is shown in table 3 (the expected frequencies range from 2 to 22). An approximate SE for each expected frequency is given by the square root of that frequency. Because of the approximations involved in deducing both the expected frequencies and the associated SEs, sequences have been classified as significantly underrepresented when the observed frequency was more than 3 SE less than the expected value.

Type I sites have the form of two tri- or tetrancleotides flanking a 6–8-base undefined spacer (see table 1). Expected frequencies for these sites were derived from the product of the frequencies of the two flanking sites.

Note that where a genome comprises regions with different base composition, these expected frequencies should be calculated separately for each of those regions. Lambda, which can be divided into two regions with quite different (G+C) content, has been treated in this way. (All analyses were performed on a DEC 20/60 computer using programs written in FORTRAN 77.)

Results

The occurrences of type I and type III restriction sites in phage genomes are presented in table 1. There is no clear pattern of avoidance of these sites. From table 3 it can be seen that in the complete genome of T7 certain 6-base palindromes—and particularly those that are recognized by restriction enzymes isolated from potential hosts of T7 (i.e., Enterobacteriaceae)—are underrepresented. Table 4 summarizes the frequencies of 6-base palindromes in all the genomes examined. φX174 exhibits a

<table>
<thead>
<tr>
<th>4-BASE CORE</th>
<th>FREQUENCY OF SURROUNDING BASES</th>
<th>Observed</th>
<th></th>
<th></th>
<th></th>
<th>Expected</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AATT</td>
<td>A-T C-G G-C T-A</td>
<td>3.5 9.1</td>
<td>9.9</td>
<td>8.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGCT</td>
<td>19 1 1 13</td>
<td>14.6 10.7</td>
<td>11.8</td>
<td>11.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GGCC</td>
<td>10.0 4.1 5.7 6.7</td>
<td>3.5 20.6</td>
<td>12.6</td>
<td>13.9</td>
<td>6.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATG</td>
<td>8.8 12.3 8.3 16.2</td>
<td>8.0 13 14.6</td>
<td>10.7</td>
<td>11.8</td>
<td>11.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCGG</td>
<td>9.9 2.0 3.6 6.6</td>
<td>8.0 13 14.6</td>
<td>10.7</td>
<td>11.8</td>
<td>11.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CGGC</td>
<td>10.0 4.1 5.7 6.7</td>
<td>3.5 20.6</td>
<td>12.6</td>
<td>13.9</td>
<td>6.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTAC</td>
<td>6.0 5.3 6.2 6.8</td>
<td>8.0 13 14.6</td>
<td>10.7</td>
<td>11.8</td>
<td>11.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTAA</td>
<td>8.0 13 14.6</td>
<td>10.7</td>
<td>11.8</td>
<td>11.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.—Letters at the top of table indicate the two bases surrounding the 4-base core given at the left. Expected frequencies were derived from observed trinucleotide frequencies (see text).

* Sequence recognized by a restriction enzyme from the Enterobacteriaceae.

b Sequence recognized by a restriction enzyme from a Bacillus species.
Table 4  
Frequencies of 6-Base Palindromes in Small Genomes

<table>
<thead>
<tr>
<th>Palindromes (N*)</th>
<th>All (64)</th>
<th>Sites (53)</th>
<th>Other (11)</th>
<th>Enterobacteriaceae (21)</th>
<th>Other (43)</th>
<th>Bacillus Species (24)</th>
<th>Other (40)</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA phages:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T7</td>
<td>263</td>
<td>185</td>
<td>78</td>
<td>22</td>
<td>241</td>
<td>41</td>
<td>222</td>
<td>39,936</td>
</tr>
<tr>
<td>φX174</td>
<td>43</td>
<td>27</td>
<td>16</td>
<td>7</td>
<td>36</td>
<td>15</td>
<td>28</td>
<td>5,386</td>
</tr>
<tr>
<td>G4</td>
<td>55</td>
<td>41</td>
<td>14</td>
<td>15</td>
<td>40</td>
<td>17</td>
<td>38</td>
<td>5,577</td>
</tr>
<tr>
<td>M13</td>
<td>0.49d</td>
<td>0.47d</td>
<td>0.54</td>
<td>0.16d</td>
<td>0.59d</td>
<td>0.28d</td>
<td>0.55d</td>
<td>6,407</td>
</tr>
<tr>
<td>fd</td>
<td>0.48d</td>
<td>0.41d</td>
<td>0.60</td>
<td>0.04d</td>
<td>0.61d</td>
<td>0.25d</td>
<td>0.55d</td>
<td>6,408</td>
</tr>
<tr>
<td>fl</td>
<td>0.45d</td>
<td>0.39d</td>
<td>0.54</td>
<td>0.08d</td>
<td>0.56d</td>
<td>0.16d</td>
<td>0.53d</td>
<td>6,407</td>
</tr>
<tr>
<td>IKe</td>
<td>0.62d</td>
<td>0.58d</td>
<td>0.69</td>
<td>0.26d</td>
<td>0.72</td>
<td>0.51</td>
<td>0.65d</td>
<td>6,883</td>
</tr>
<tr>
<td>Lambda</td>
<td>0.54d</td>
<td>0.52d</td>
<td>0.66d</td>
<td>0.44d</td>
<td>0.60d</td>
<td>0.51d</td>
<td>0.56d</td>
<td>48,502</td>
</tr>
<tr>
<td>φ29</td>
<td>0.79</td>
<td>0.56d</td>
<td>1.26</td>
<td>0.51</td>
<td>0.84</td>
<td>0.23d</td>
<td>0.92</td>
<td>5,708</td>
</tr>
<tr>
<td>RNA phage:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS2</td>
<td>60</td>
<td>53</td>
<td>7</td>
<td>22</td>
<td>38</td>
<td>24</td>
<td>36</td>
<td>3,569</td>
</tr>
<tr>
<td>DNA viruses:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EBV</td>
<td>2684</td>
<td>2489</td>
<td>195</td>
<td>1705</td>
<td>979</td>
<td>1746</td>
<td>938</td>
<td>172,282</td>
</tr>
<tr>
<td>AD2</td>
<td>631</td>
<td>555</td>
<td>76</td>
<td>293</td>
<td>338</td>
<td>313</td>
<td>318</td>
<td>35,937</td>
</tr>
<tr>
<td>Papilloma</td>
<td>117</td>
<td>73</td>
<td>44</td>
<td>23</td>
<td>94</td>
<td>24</td>
<td>93</td>
<td>7,811</td>
</tr>
<tr>
<td>SV40</td>
<td>88</td>
<td>67</td>
<td>21</td>
<td>29</td>
<td>59</td>
<td>18</td>
<td>70</td>
<td>5,242</td>
</tr>
<tr>
<td>Mitochondrial</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>genomes:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mt cow</td>
<td>250</td>
<td>156</td>
<td>94</td>
<td>40</td>
<td>210</td>
<td>41</td>
<td>209</td>
<td>16,338</td>
</tr>
<tr>
<td>mt man</td>
<td>211</td>
<td>145</td>
<td>66</td>
<td>63</td>
<td>148</td>
<td>50</td>
<td>161</td>
<td>16,569</td>
</tr>
<tr>
<td>mt mouse</td>
<td>295</td>
<td>146</td>
<td>49</td>
<td>45</td>
<td>250</td>
<td>46</td>
<td>249</td>
<td>16,295</td>
</tr>
</tbody>
</table>

Note.—Palindromes have been grouped according to whether or not they are recognized by any known restriction enzyme (Sites), enzymes derived from Enterobacteriaceae, or enzymes derived from Bacillus species.

*a Number of 6-base palindromes in each category.
*b Observed frequency.
*c Observed/expected.
*d Significantly less than 1.00.

pattern of site avoidance similar to that of T7, as do the four filamentous phages. Although in several of these phages there appears to be avoidance of Bacillus enzyme sites, it can be seen from table 2 that there is considerable overlap between the sets of
sites recognized by enterobacterial and *Bacillus* enzymes. In φ29, enterobacterial enzyme sites are not so clearly avoided, but those that would be recognized by endonucleases from *Bacillus* species are. In fact there are no occurrences of the 14 6-base palindromes that would be recognized by *B. subtilis* enzymes (six would be expected, but this is not formally significant, by the criteria outlined above). In G4 6-base palindromes in general are not significantly underrepresented, nor is there any particular avoidance of those that are restriction sites. The data for lambda is intermediate between that for G4 and that for the other DNA phages. Six-base palindromes are significantly underrepresented, but the effect is not much more pronounced for enterobacterial restriction sites than for any others. There are 145 sites for “host” restriction enzymes, and none of the recognition sequences is completely absent. Six-base palindromes are not underrepresented in either MS2 or in any of the eukaryote viruses or mitochondrial genomes examined here.

**Discussion**

Restriction by host endonucleases exerts selection pressure on bacteriophages. The first T7 gene to be expressed after infection, the 0.3 gene, encodes a protein that inhibits the action of the type I restriction enzymes EcoB and EcoK (Studier 1975). Thus selection against the recognition sites of these type I enzymes would not be expected, and indeed these sites are found in the T7 genome. A similar gene is found in the related phage T3, and the products of each are also effective against the SA and SB restriction systems of *Salmonella typhimurium* (Kruger et al. 1983). This gene product is not effective against type II restriction enzymes, and indeed no active defense against these enzymes is known in T7. However, natural selection against phages with the recognition sites for these enzymes in their DNA seems to have produced a passive defense—i.e., avoidance of these particular short sequences. Type III restriction enzymes have properties intermediate between those of type I and type II (Yuan 1981), and it is not clear whether the 0.3 gene product is an effective defense. From the presence of type III recognition sites within the T7 genome it must be concluded either that T7 has some defense against these enzymes or that ancestors of T7 have had no contact with bacterial strains encoding them.

That the underrepresentation of certain 6-base palindromes in DNA coliphages results from selection against phages containing restriction sites is strongly supported by two points. One, the particular sequences avoided correspond closely to those recognized by restriction enzymes derived from *E. coli* or a related bacterium. Two, the RNA coliphage MS2 (recall that restriction enzymes do not cleave RNA) and other small DNA genomes—i.e., eukaryotic viruses and mammalian mitochondrial molecules—do not show this avoidance.

Of the other coliphages, φX174 and the four related filamentous phages (M13, fd, f1, and lKe) also exhibit avoidance of these same sites. (Note that M13 and fd, and to a lesser extent f1, are so closely related that they perhaps do not represent “independent” data.) φX174, G4, and the filamentous phages are all single stranded but pass through double-stranded replicative forms when they could be susceptible to the action of restriction endonucleases. There is also evidence that cleavage can occur at duplex regions in single-stranded φX174 (Blakesley et al. 1977). The *Bacillus subtilis* phage φ29 shows a significant avoidance of the recognition sites of enzymes isolated from *Bacillus* species. These findings are so intuitively reasonable that it is perhaps of more interest to ask why G4 and, to a lesser extent, lambda do not conform to this pattern.
It is known that lambda is heavily restricted (Court and Oppenheim 1983), and so it is not easy to see why this does not impose selection pressure for the avoidance of restriction sites. Lambda does differ from the other phage in that it is temperate (the others are lytic), and while integrated into the *E. coli* genome it is not expected to suffer restriction.

G4 was originally isolated in a search for φX174-like phages (Godson 1974), and there is ~65% identity between the two (Godson et al. 1978). Although cleavage of these two phages produces quite different restriction fragments (Godson and Roberts 1976), it is surprising that such a large difference exists between the two in frequency of restriction sites. Perhaps the G4 double-stranded replicative form is less vulnerable. Alternatively, G4 may be adapted to a different host.

Some sequences other than those known to be potential host restriction sites are also underrepresented in the phages examined. The *E. coli* dam methylation site (GATC) is rare in the genomes of enterobacteriophages (McClelland 1985). It is also likely that more restriction enzymes will be isolated, with specificities new to the species from which they are derived (some predictions might be made from the sites widely avoided).

Finally, the question arises, Why are there any sites within the genome of, for example, T7 that can be recognized by an *E. coli* restriction enzyme? One answer may lie in the (largely unknown) population structure of bacteria and bacteriophages. It is not clear (1) how many of the genes encoding the restriction enzymes isolated in the laboratory from species of Enterobacteriaceae might be found in any one wild population of bacteria; (2) how often these genes, perhaps on plasmids, might transfer between populations; and (3) how often phages transfer between populations of bacteria. It is concluded from the data presented here that not all enzymes are produced by all populations and that the different phages have not been exposed to certain of these enzymes in their recent history.

Acknowledgments

I thank M. McClelland for making available his results prior to their publication, D. J. McConnell for stimulating discussion, and R. Grantham for his comments on the manuscript.

LITERATURE CITED


STUDIER, F. W. 1975. Gene 0.3 of bacteriophage T7 acts to overcome the DNA restriction system of the host. J. Mol. Biol. 94:283–295.


MASATOSHI NEI, reviewing editor

Received July 17, 1985; revision received September 4, 1985.
The number of charge changes relative to total amino acid replacements for each of seven protein sequences (cytochrome c, hemoglobin α, hemoglobin β, myoglobin, insulin, and fibrinopeptides A and B) has been studied. This number was compared with the expected value obtained under the assumption of random nucleotide substitution. The results obtained indicate that four proteins—hemoglobin α, hemoglobin β, myoglobin, and insulin—are accumulating charge changes at rates slower than those predicted by a model of random substitution. Cytochrome c and fibrinopeptides A and B are accumulating charge changes at rates similar to those predicted by a random model.

Introduction

Application of the techniques of molecular biology has provided both new insights into evolutionary processes and the development of new methods in the reconstruction of evolutionary history (e.g., Ayala 1976; Nei and Koehn 1983). Comparison of the amino acid sequences of proteins from extant species provides detailed information regarding evolutionary patterns and rates. Investigation of the rate of amino acid replacements has led to the observation that point mutations appear to occur and be fixed in the genome at reasonably regular rates (e.g., Nei 1975). Fundamental to an understanding of protein evolution is the elucidation of the underlying mechanisms responsible for the accumulation of amino acid replacements. Specifically, assuming that mutations occur at random at the nucleotide level, exactly how does natural selection edit these alterations to permit an amino acid to replace a previously acceptable codon? Can generalizations be made (e.g., replacement by chemically similar residues [Grantham 1974; Marshall and Brown 1975]), or is each case unique, being dependent on the role that a particular amino acid plays in the action of that molecule?

The intent of this study is to investigate one particular constraint on amino acid replacement—namely, as a subset of amino acids diverge in a pair of homologous polypeptides, are the four charged residues replaced at random? In this study, we compared seven protein sequences (cytochrome c, hemoglobin α, hemoglobin β, myoglobin, insulin, and fibrinopeptides A and B) position by position against homologous sequences from a variety of species, and the number of charge changes that occurred relative to total amino acid replacements were calculated. Comparisons were made between these observed values and expected values calculated for each protein assuming random substitution. Ionic associations between amino acid side chains are known to be important components of protein structure, as well as to form contacts in mul-

1. Key words: molecular evolution, charge change, cytochrome c, hemoglobins, myoglobin, insulin, fibrinopeptides.

Address for correspondence and reprints: Dr. Glenys Thomson, Department of Genetics, Mulford Hall, University of California, Berkeley, California 94720.

© 1986 by The University of Chicago. All rights reserved.
0737-4038/86/0301-3103$02.00

84
timeric systems, and could well be a major selective component. Charge changes in protein evolution could be important to electrophoresis-based research that depends heavily on charge variation for identification of allelic variants.

**Methods**

Single amino acids have three points that may ionize to give a charged molecule: the $\alpha$-carboxyl group (ionized negatively at pH values above the 2–3 range), the $\alpha$-amino group (ionized positively at pH values below 8.5), and the side group. Peptide bonding stabilizes all but one each of the $\alpha$-amino and $\alpha$-carboxyl groups, and these two protein termini have been ignored, since they will be constant across all species. At physiological (and most electrophoretic) pH values, four side chains will be ionized: arginine and lysine will be positively charged, and aspartic acid and glutamic acid will be negatively charged. In summary, each amino acid position in a protein will be defined as being either neutral (16 amino acids), +1 (2 amino acids), or −1 (2 amino acids).

Given this protocol, and using the genetic code, we find that the calculation of the probability that mutations will lead to a charge change for a given amino acid is straightforward. The expectation that charge change will occur in the four charged amino acids is high, whereas that for neutral amino acids is low. Assuming a uniform amino acid composition, we find that the probability that an amino acid replacement will involve a charge change is 0.327 (Nei 1975, p. 25). However, given the fact that proteins seldom have equal proportions of all amino acids, this theoretical expectation of charge change must be calculated on a per-protein basis. For example, human cytochrome c, with 29% of its amino acids charged, would be expected to change charge at a higher rate than human insulin, which contains only 12% charged amino acids.

The seven proteins included in this study were cytochrome c, hemoglobin $\alpha$, hemoglobin $\beta$, myoglobin, insulin, fibrinopeptide A, and fibrinopeptide B. The species from which the amino acid sequences of these proteins were obtained are listed in table 1. The data for all proteins but myoglobin were from Dayhoff (1972, 1973, 1976, 1978), whereas the myoglobin data came from Romero-Herrera et al. (1978).

The method of analysis was as follows. First, we analyzed the observed data. For each protein a position-by-position comparison between each pair-wise combination of sequences was made to determine the total number of amino acid replacements. All possible pairwise comparisons between species were made. The data as presented in Dayhoff are such that, with the use of well-placed gaps, even for very different species a majority of the aligned amino acids are identical. Only positions that had one of the 20 common amino acids on both sequences were compared; positions with a gap or an atypical or indeterminate amino acid (usually indistinguishable between aspartic acid and asparagine or between glutamic acid and glutamine) were exempted from further consideration. A tally was kept of the number of amino acid replacements and of the charge differences for each pair. The results of all pairwise comparisons formed a bivariate sample from which we calculated, for each of the seven proteins studied, the slope forced through the origin, SE, and confidence intervals for the ratio of the amount of charge change (dependent variable) to total amino acid replacement (independent variable) (Nie et al. 1975).

Because proteins differ widely in their amino acid compositions, it is conceivable that each of the seven proteins analyzed could have a widely different average prob-
<table>
<thead>
<tr>
<th>SPECIES</th>
<th>Cytochrome c</th>
<th>Hb α</th>
<th>Hb β</th>
<th>Myoglobin</th>
<th>Insulin</th>
<th>Fibrinopeptide A</th>
<th>Fibrinopeptide B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.349</td>
<td>0.200</td>
<td>0.212</td>
<td>0.226</td>
<td>0.150</td>
<td>0.309</td>
<td>0.597</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0.309</td>
<td>0.597</td>
</tr>
<tr>
<td>Gorilla</td>
<td>0.357</td>
<td>0.165</td>
<td>0.209</td>
<td>0.222</td>
<td>0.309</td>
<td>0.597</td>
<td>0.597</td>
</tr>
<tr>
<td>Gibbon</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Baboon</td>
<td>...</td>
<td>0.162</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Irus macaque</td>
<td>...</td>
<td>0.229</td>
<td>0.211</td>
<td>0.163</td>
<td>0.316</td>
<td>0.424</td>
<td>0.424</td>
</tr>
<tr>
<td>Orangutan</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0.309</td>
<td>0.351</td>
</tr>
<tr>
<td>Rhesus monkey</td>
<td>...</td>
<td>0.206</td>
<td>0.204</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Squirrel monkey</td>
<td>...</td>
<td>...</td>
<td>0.199</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Woolly monkey</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0.262</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Common marmoset</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0.247</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Thick-tailed galago</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0.210</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Potto</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0.205</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Slow loris</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0.239</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Sportive lemur</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0.204</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Dog</td>
<td>0.385</td>
<td>0.177</td>
<td>0.197</td>
<td>0.169</td>
<td>...</td>
<td>0.463</td>
<td>0.521</td>
</tr>
<tr>
<td>Hunting dog</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0.179</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Cat</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0.250</td>
<td>0.519</td>
<td>...</td>
</tr>
<tr>
<td>Horse</td>
<td>0.398</td>
<td>0.228</td>
<td>0.240</td>
<td>0.202</td>
<td>0.154</td>
<td>0.390</td>
<td>0.543</td>
</tr>
<tr>
<td>Donkey</td>
<td>0.415</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0.407</td>
<td>0.543</td>
</tr>
<tr>
<td>Llama</td>
<td>...</td>
<td>0.233</td>
<td>0.249</td>
<td>...</td>
<td>...</td>
<td>0.436</td>
<td>0.607</td>
</tr>
<tr>
<td>Pig</td>
<td>0.392</td>
<td>0.299</td>
<td>0.212</td>
<td>...</td>
<td>0.156</td>
<td>0.324</td>
<td>0.475</td>
</tr>
<tr>
<td>Bovine</td>
<td>...</td>
<td>0.273</td>
<td>0.290</td>
<td>0.231</td>
<td>0.146</td>
<td>0.306</td>
<td>0.447</td>
</tr>
<tr>
<td>Sheep</td>
<td>...</td>
<td>0.342</td>
<td>0.171</td>
<td>0.250</td>
<td>0.145</td>
<td>0.340</td>
<td>0.420</td>
</tr>
<tr>
<td>Goat</td>
<td>...</td>
<td>0.241</td>
<td>0.175</td>
<td>...</td>
<td>...</td>
<td>0.340</td>
<td>0.420</td>
</tr>
<tr>
<td>Common treeshrew</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0.164</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>European hedgehog</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0.306</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Badger</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0.208</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Mouse</td>
<td>0.361</td>
<td>0.195</td>
<td>0.214</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>Rat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guinea pig</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rabbit</td>
<td>0.390</td>
<td>0.205</td>
<td>0.205</td>
<td></td>
<td></td>
<td>0.290</td>
<td></td>
</tr>
<tr>
<td>Grey kangaroo</td>
<td>0.320</td>
<td>0.253</td>
<td>0.219</td>
<td>0.250</td>
<td></td>
<td>0.583</td>
<td>0.551</td>
</tr>
<tr>
<td>Opossum (North America)</td>
<td></td>
<td></td>
<td></td>
<td>0.165</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harbor seal</td>
<td></td>
<td></td>
<td></td>
<td>0.172</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sperm whale</td>
<td></td>
<td></td>
<td></td>
<td>0.152</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dolphin</td>
<td></td>
<td></td>
<td></td>
<td>0.199</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forpoise</td>
<td></td>
<td></td>
<td></td>
<td>0.211</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chicken</td>
<td>0.359</td>
<td>0.178</td>
<td>0.227</td>
<td>0.208</td>
<td>0.133</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pekin duck</td>
<td>0.379</td>
<td></td>
<td></td>
<td></td>
<td>0.213</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Figeon</td>
<td>0.366</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>King penguin</td>
<td>0.356</td>
<td></td>
<td></td>
<td>0.189</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bullfrog</td>
<td>0.364</td>
<td></td>
<td>0.330</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rattlesnake</td>
<td>0.403</td>
<td></td>
<td></td>
<td>0.207</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lizard</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.758</td>
<td></td>
</tr>
<tr>
<td>Newt</td>
<td></td>
<td>0.368</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carp</td>
<td>0.270</td>
<td>0.178</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tuna</td>
<td>0.362</td>
<td></td>
<td></td>
<td>0.239</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bonita</td>
<td>0.348</td>
<td></td>
<td></td>
<td>0.149</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pacific lamprey</td>
<td>0.372</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fruit fly</td>
<td>0.311</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Screwworm fly</td>
<td>0.329</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silkworm moth</td>
<td>0.328</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hornworm moth</td>
<td>0.334</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wheat</td>
<td>0.377</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sunflower</td>
<td>0.362</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Castor bean</td>
<td>0.335</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mung bean</td>
<td>0.346</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sesame</td>
<td>0.349</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Debaromyces kloeckeri</td>
<td>0.463</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Candida kruzei</td>
<td>0.394</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baker's yeast</td>
<td>0.415</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neurospora crassa</td>
<td>0.323</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total no. of unique sequences</td>
<td>32</td>
<td>18</td>
<td>18</td>
<td>29</td>
<td>12</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>Mean ± SD</td>
<td>0.361 ± 0.038</td>
<td>0.232 ± 0.057</td>
<td>0.220 ± 0.038</td>
<td>0.209 ± 0.036</td>
<td>0.179 ± 0.041</td>
<td>0.379 ± 0.121</td>
<td>0.500 ± 0.077</td>
</tr>
</tbody>
</table>
ability of charge replacement. Our next step was to calculate the theoretical value for each protein, assuming random substitution. Two methods were used to calculate the expected relationship between charge change and amino acid replacement. Both methods used as a starting point a consensus protein sequence determined from all species for which data were available. In order to obtain a range of values, initial sequences were also taken from the three species with the highest and the three species with the lowest total charge for each protein.

In the first method (A), for each protein the proportion of amino acids resulting in a charge change was calculated theoretically, using the genetic code. In the second method (B), for each protein the divergence of two homologous sequences to the observed level of amino acid divergence was simulated. In this case, an initial sequence was generated from the overall distribution of all species. (Later, in an effort to determine a range of values, sequences were generated from distributions taken from the three species with the highest and the three species with the lowest total charge.) This sequence was copied to provide two identical initial sequences. Independent nucleotide hits were simulated in the following manner: one of the two sequences was randomly chosen; one of the amino acid positions within this sequence was chosen at random; one of the three nucleotide positions within this amino acid was chosen at random; and then one of the three nucleotides was chosen at random to substitute for this nucleotide. Generation of a stop codon necessitated an alternative nucleotide. The data from the simulation were subject to regression analysis in the same manner as the empirical data. The simulation was performed 100 times per protein.

Results

To more easily detect possible heterogeneity in the data, we also calculated for each species and protein the proportion of total charge changes among total amino acid replacements in the comparison of that species versus all others (see table 1). The average of these values for a given protein is given at the end of table 1. The distribution of these values was not statistically different from a normal distribution, except in the case of fibrinopeptide A. For all proteins except fibrinopeptide B and insulin, there are outliers, that is, values that are more than twice the SD from the mean. There is only one outlier that is less than the mean, the comparison of carp to all other species for cytochrome c (0.270). The other outliers are *Debaryomyces kloeckeri* for cytochrome c (0.463), newt for hemoglobin α (0.368), bullfrog for hemoglobin β (0.330), hedgehog for myoglobin (0.306), and lizard for fibrinopeptide A (0.758).

For the set of all pairwise comparisons for a given protein, a regression analysis was performed, and the observed proportion of charge changes to total amino acid replacements (observed slope, forced through the origin) for each of the seven proteins, along with SE and confidence intervals, is given in table 2. These values vary from a low of 0.205 for myoglobin to a high of 0.572 for fibrinopeptide B. Note that for each protein the mean values in table 1 and the slopes in table 2 are very similar, though a direct comparison cannot be made since they were obtained by different methods.

The plots of all pairwise comparisons, the fitted regression line (forced through the origin), and the expected line assuming random substitutions are given for cytochrome c and myoglobin in figures 1A and 1B, respectively.

The ratios of charge changes to amino acid replacements obtained by method B are very close to the theoretical values obtained by method A (see table 3). These
Table 2
Results of the Regression Analysis of Charge Changes to Total Amino Acid Replacements for Each of the Seven Proteins Studied in the Species Listed in Table 1

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sample Size</th>
<th>Observed Slope</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytochrome c</td>
<td>496</td>
<td>0.370 ± 0.003</td>
<td>0.364, 0.376</td>
</tr>
<tr>
<td>Hemoglobin α</td>
<td>153</td>
<td>0.244 ± 0.008</td>
<td>0.228, 0.260</td>
</tr>
<tr>
<td>Hemoglobin β</td>
<td>153</td>
<td>0.251 ± 0.006</td>
<td>0.239, 0.263</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>406</td>
<td>0.205 ± 0.003</td>
<td>0.199, 0.211</td>
</tr>
<tr>
<td>Insulin</td>
<td>66</td>
<td>0.212 ± 0.008</td>
<td>0.196, 0.228</td>
</tr>
<tr>
<td>Fibrinopeptide A</td>
<td>190</td>
<td>0.378 ± 0.015</td>
<td>0.348, 0.408</td>
</tr>
<tr>
<td>Fibrinopeptide B</td>
<td>153</td>
<td>0.572 ± 0.014</td>
<td>0.544, 0.600</td>
</tr>
</tbody>
</table>

* No. of pairwise comparisons.

values ranged from a low of 0.295 for insulin to a high of 0.520 for fibrinopeptide B. Simulations and theoretical calculations from low-charge and high-charge species broadened the range of slope values somewhat, to the greatest degree in the case of fibrinopeptide B.

For the comparison of observed and expected proportions of charge changes, the simulation results, rather than the theoretical results, are used for the expected values because they allow an estimate of the SE of the expected values. The observed and expected slopes from the regression analyses (ratio of charge changes to total replacements), as well as the SEs of these estimates, are summarized in table 4. This provides a test to detect differences between observed and expected values.

Five of the proteins—cytochrome c, hemoglobin α, hemoglobin β, myoglobin, and insulin—have slopes that are significantly less than those of the simulated values. These range from myoglobin, in which the observed slope is only 51% of the simulated value, to cytochrome c, which is 92% of the simulated value. However, if we correct the probability levels for the number of comparisons that have been made (i.e., seven proteins), the observed cytochrome c results are not significantly different from those predicted by a model of random substitution. The observed results for fibrinopeptide A are not significantly different from the simulated expectations for it, whereas for fibrinopeptide B the observed slope is significantly greater than that for the simulated results. However, the average slope for fibrinopeptide B, calculated from a priori expectations, is 0.520, which is not significantly different from the observed slope of 0.572.

Discussion

Four proteins—hemoglobin α, hemoglobin β, myoglobin, and insulin—of the seven studied appear to be accumulating charge changes more slowly than they would be expected to do at random, suggesting that there may be strong selection against charge changes in these molecules. On the other hand, cytochrome c and fibrinopeptides A and B have rates of accumulation of charge changes similar to those predicted by a random model. These results are not contradicted by observations of particular species that exhibit either more or fewer charge changes than predicted by the genetic code (Hewett-Emmett et al. 1976). Outliers have been identified in our study.
FIG. 1.—Plots of all pairwise comparisons of the proportion of charged amino acid changes (y-axis) vs. the proportion of total amino acid replacements (x-axis), the fitted regression line forced through the origin (solid line), and the expected line assuming random substitution (dotted line) for cytochrome c (A) and myoglobin (B). ○ = 1; ● = 2; ▲ = 3; ▼ = 4; ■ = 5; □ = 6; × = 7; ◄ = 8; and ▼ = 9.
Table 3
Proportion of Amino Acid Replacements Resulting in Charge Changes for the Protein Examined

<table>
<thead>
<tr>
<th>PROTEIN AND SEQUENCE(S)</th>
<th>CHARGE CHANGES/TOTAL REPLACEMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A Priori Expectation (Method A)</td>
</tr>
<tr>
<td>Cytochrome c:</td>
<td></td>
</tr>
<tr>
<td>High-charge sequence</td>
<td>0.401</td>
</tr>
<tr>
<td>All sequences</td>
<td>0.391</td>
</tr>
<tr>
<td>Low-charge sequence</td>
<td>0.375</td>
</tr>
<tr>
<td>Hemoglobin a:</td>
<td></td>
</tr>
<tr>
<td>High-charge sequence</td>
<td>0.329</td>
</tr>
<tr>
<td>All sequences</td>
<td>0.337</td>
</tr>
<tr>
<td>Low-charge sequence</td>
<td>0.351</td>
</tr>
<tr>
<td>Hemoglobin β:</td>
<td></td>
</tr>
<tr>
<td>High-charge sequence</td>
<td>0.341</td>
</tr>
<tr>
<td>All sequences</td>
<td>0.339</td>
</tr>
<tr>
<td>Low-charge sequence</td>
<td>0.351</td>
</tr>
<tr>
<td>Myoglobin:</td>
<td></td>
</tr>
<tr>
<td>High-charge sequence</td>
<td>0.313</td>
</tr>
<tr>
<td>All sequences</td>
<td>0.305</td>
</tr>
<tr>
<td>Low-charge sequence</td>
<td>0.305</td>
</tr>
<tr>
<td>Insulin:</td>
<td></td>
</tr>
<tr>
<td>High-charge sequence</td>
<td>0.340</td>
</tr>
<tr>
<td>All sequences</td>
<td>0.304</td>
</tr>
<tr>
<td>Low-charge sequence</td>
<td>0.295</td>
</tr>
<tr>
<td>Fibrinopeptide A:</td>
<td></td>
</tr>
<tr>
<td>High-charge sequence</td>
<td>0.432</td>
</tr>
<tr>
<td>All sequences</td>
<td>0.439</td>
</tr>
<tr>
<td>Low-charge sequence</td>
<td>0.434</td>
</tr>
<tr>
<td>Fibrinopeptide B:</td>
<td></td>
</tr>
<tr>
<td>High-charge sequence</td>
<td>0.422</td>
</tr>
<tr>
<td>All sequences</td>
<td>0.434</td>
</tr>
<tr>
<td>Low-charge sequence</td>
<td>0.434</td>
</tr>
</tbody>
</table>

NOTE.—See text for details of the two methods used.

The rates of amino acid substitution for fibrinopeptides A and B are among the fastest known (Dayhoff 1972). Fibrinopeptides are extremely short molecules, with approximately 16 residues for A and approximately 13 for B. It is difficult to obtain alignments of fibrinopeptides from different species. The shortness and rapid evolutionary rate of these proteins mean that the amino acid composition of the fibrinopeptides will show a large variance and be particularly prone to large stochastic effects. This is demonstrated by the wide range of the ratios of expected rate of charge change to total amino acid replacements calculated from all fibrinopeptide B sequences (0.520) compared with those calculated for the three low-charge (0.569) and the three high-charge (0.434) sequences (see table 3). Such a wide range is not seen in the other proteins. The fact that the observed rate of change in fibrinopeptide A matches the expected rate under a random model but that the fibrinopeptide B rate of accumulation...
Table 4
Comparison of Mean ± SE Values for Observed and Expected Proportions of Charge Replacements to Total Amino Acid Changes (Slope)

<table>
<thead>
<tr>
<th>Protein</th>
<th>Observed</th>
<th>Expected</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytochrome c</td>
<td>0.370 ± 0.003</td>
<td>0.403 ± 0.014 (18)</td>
<td>&lt;0.025</td>
</tr>
<tr>
<td>Hemoglobin a</td>
<td>0.244 ± 0.008</td>
<td>0.335 ± 0.013 (15)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Hemoglobin β</td>
<td>0.251 ± 0.006</td>
<td>0.346 ± 0.014 (22)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>0.205 ± 0.003</td>
<td>0.403 ± 0.014 (19)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Insulin</td>
<td>0.212 ± 0.008</td>
<td>0.312 ± 0.019 (40)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Fibrinopeptide A</td>
<td>0.378 ± 0.015</td>
<td>0.447 ± 0.051 (39)</td>
<td>NS</td>
</tr>
<tr>
<td>Fibrinopeptide B</td>
<td>0.572 ± 0.014</td>
<td>0.449 ± 0.039 (36)</td>
<td>&lt;0.005</td>
</tr>
</tbody>
</table>

NOTE.—The numbers in parentheses are the sample size (number of 100 simulation sets) from which the SE was calculated. NS = not significant.

*a See table 2.

*b See table 3.

of charge changes is higher than the simulation (the a priori expectations are not different) could result merely from such stochastic effects.

The regression analysis that we have performed to calculate the ratio of observed charge changes to total amino acid charges for each protein type is not strictly valid. The regression procedure involves the use of data points from comparisons between pairs of species. Since all possible pairwise comparisons have been made and used in the analysis, this means that all the data points are not independent (Nei and Chakraborty 1976). However, because the number of comparisons is so large for each protein—that is, \( n(n - 1)/2 \), where \( n \) is the number of species sequences for a particular protein—this effect should be small. Another potential bias occurs because the ancestral sequences are the same for compared lineages. However, this bias is the same for both the observed and the simulated values. Except in the case of fibrinopeptide B, the three low-charge and three high-charge sequences’ simulated and theoretical values did not differ greatly from the values calculated using all sequences, a result that further validated this approach.

In our simulations, we also kept track of total protein charge in our diverging lineages. We found that when total protein charge was compared across several species there was a fairly narrow range of total charge for all the homologous sequences, even when charge changes accrue at random or close to random rates. This is because the total protein charge of a diverging sequence is highly dependent on the starting value and hovers around that value for a very long time. It has been noted that a strong predilection exists for maintenance of overall protein polarity at a given level throughout all species (Vogel 1971; Vogel and Zuckermandl 1971). Our simulations, although not in any way disproving the possibility of selection for total constant charge, indicate that this observation is also consistent for a model with purely random accumulation of charge changes.

All theoretical expectations in this study have been calculated under the assumption of purely random nucleotide substitution, in terms of both the type of substitution that will take place and the site of the substitution. Experimental results suggest several nonrandom components in substitution. Li et al. (1984) found that
transitions occur almost twice as often as expected under random mutation. Also, although mutations most likely occur at random with respect to triplet position, the same cannot be true once natural selection acts on the changes. Third-nucleotide substitutions are notable for resulting in similar amino acids. Fitch (1973) and Grunstein et al. (1976) both found marked differences in the propensity of the three positions to permanently establish a novel nucleotide. Additionally, in any protein certain amino acid positions are virtually immune from change (Fitch and Margoliash 1967), whereas others can be considered hypervariable (Fitch and Markowitz 1970). Even more complicating for any model attempting to predict amino acid replacements is the postulation of the covarion (Fitch and Markowitz 1970; Fitch 1971), which states that some relatively constant percentage of the codons of any single sequence are mutable but that as a position does mutate it acts as a pioneer by providing new areas for mutation.

As a first step in the analysis of the forces operating to determine the ratio of the rate of evolution of charge replacements to the rate of total changes in a protein, we have assumed the simplest possible model of random nucleotide substitution. For four of seven proteins examined, the ratio of charge changes to total amino acid replacements was significantly less than this random rate, suggesting that negative selection is important in reducing the rate of charge change (e.g., see Kimura 1983).

This observation appears to be consistent with the observation of Grantham (1974) that chemically similar amino acids are interchanged more easily than chemically dissimilar ones. A more detailed and realistic model than the one we have considered might specify some requirements and restrictions on allowable amino acid changes for each protein and elucidate the mechanisms responsible for the reduced rate of charge-change replacement. We note that the present analysis cannot determine whether those mutations that do become established in a population are either selectively equivalent to or different from each other.

Acknowledgments

We would like to thank S. Selvin, V. Sarich, U. Motro, W. Klitz, and E. Louis for their helpful discussions and comments on this study, O. Mayo for his comments on the manuscript, and Anne Gilligan and Irva Hertz for data analysis. Grant support from National Institutes of Health grant HD 12731 and Biomedical Research grant IS07 RR 07006 is gratefully acknowledged.

LITERATURE CITED


MASATOSHI NEI, reviewing editor

Received June 10, 1985; revision received July 16, 1985.

Traditional evolutionary biologists have a deep-seated mistrust of microbes. Microorganisms are usually responsible for causing them a certain amount of embarrassment and discomfort whilst on field trips in tropical and subtropical parts of the world, and, suffering in the confines of their tent (or, if they are lucky, in a hotel room), they usually dismiss the possibility that microbes can provide any valuable information concerning evolution. How can, they would argue, anything useful come out of the study of organisms that have neither fur nor feathers, no easily visible chromosomes, no regularly occurring sexual cycle, and, last but certainly not least, no fossil record. Many, but unfortunately not all, of the answers to these questions will be found in this book, edited by Robert Mortlock, one of the first people to realize the worth of microbial studies to evolutionary biology.

One of the fascinating aspects of microbial evolution is the incredible range of metabolic activities possessed by microbes and the seeming ease and speed with which they can acquire new ones. The prize for the most bizarre microbe must surely go to the bacterium that lives and divides at temperatures of 250°C or more, temperatures at which all self-respecting proteins become denatured and DNA double helices fall apart. The rapid acquisition of resistance or ability to degrade exotic new macromolecules in the environment is even more spectacular from an evolutionary point of view. The rapid spread of antibiotic-resistant organisms in nosocomial environments is now well documented. Less well known, but no less interesting, is the emergence of strains able to degrade such new and toxic components of our environment as PCBs, the pesticide 2,4-D, and 6-aminohexanoic acid, a by-product of the nylon industry. With the exception of the last chapter, a provocative discussion of bacterial genome evolution by Monica Riley, this book deals exclusively with this latter aspect—that is, evolution of enzymatic function—and, in being so restrictive, it must be faulted. A full 50% of the book, five chapters, is devoted to articles concerned with the metabolism of pentoses and pentitols, and four of these five chapters are written by only two authors—Mortlock himself and Brian Hartley. One glaring omission is a chapter dealing with plasmids and with lysogenic and lytic phages. The evolutionary fates of bacterial and episomal genomes, existing together within the same cell, are so intertwined that they must be ideal model systems for coevolutionary studies. A number of possible authors who could have made a worthwhile contribution in this area immediately come to mind. Plasmids are also intimately involved in the acquisition of new metabolic function, and a chapter discussing their role would have been particularly appropriate given the emphasis of this book.

Since this is really the first book to appear that is exclusively devoted to microorganisms as subjects for evolutionary studies, I would have preferred to have seen a wider treatment of the subject. Articles dealing with the aspects of microbial evolution dealt with by such workers as Daniel Hartl and Bruce Levin would have done much to widen the appeal of this book. The growing number of evolutionary biologists working with microorganisms is a testament to the many intrinsic advantages that
microbes have for studies in evolutionary biology. These workers should be heartened by the appearance of this book, and it should be read by the vast majority of evolutionary biologists not working with microorganisms. If this book does no more than stimulate other researchers to enter the field and provoke the publication of another volume in this area it will have been a great success.

JULIAN ADAMS
University of Michigan
Books Received


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 without the written permission of the copyright owner. The journal will not be responsible for the loss of manuscripts at any time.

Publication is taken 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.
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 (analog) 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 4 3/4 inches wide by 7 3/4 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

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

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

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
Subscribe to

BioScience
The biologist's monthly magazine

For professionals and students in every field of biology: Research and policy articles, features, news updates, book reviews, announcements and professional opportunities, letters, and a look at today's computer technologies and new products.

Join AIBS and start your subscription today.

Name ____________________________________________________________

Address __________________________________________________________

Charge my [ ] Credit Card [ ] Acct. # ________________________________

Signature _________________________________________________________

Rates: [ ] Individual @ $37.50 per year, [ ] Family @ $52.00 per year, [ ] Sustaining @ $55.00 per year (These include $23.00 per year for BioScience), [ ] Student @ $19.50 per year, [ ] Emeritus @ $19.50 per year (These include $12.00 per year for BioScience)

Please make checks payable to AIBS. Remit US dollars only. Mail to: AIBS, P.O. Box 27417, Central Station, Washington, DC 20038-7417.

American Institute of Biological Sciences
A critical publication for everyone working in molecular biology . . .

Now in its fifth year of publication

DNA

A Journal of Molecular Biology

Editors:
Walter L. Miller, M.D.
and
James Roberts, Ph.D.

FROM LATEST ISSUES

Expression of Rat NADPH—Cytochrome P450 Reductase cDNA in Saccharomyces cerevisiae.
By H. Murakami, Y. Yabusaki and H. Okawa

A General Method for Retrieving the Components of a Genetically Engineered Fusion Protein.
By P.K. Szoka, A.W. Schreiber, and H. Chan

Synthesis of Bovine Prolactin in Escherichia coli.
By D.N. Luuk, J.K. Ngoe, F.M. Rottman and M. Smith

Isolation and Characterization of the α1-Antitrypsin Gene of Mice.
By K.S. Krauter, B.A. Citron, M.-T. Hsu, D. Powell and J.E. Darnell, J.R.

Construction of a Systematic Set of sRNA Mutants by Ligation of Synthetic Oligonucleotides into Defined Single-Stranded Gaps.
By S.W. Cline, M. Yarus and P. Wier

Binding of the Human Glucocorticoid Receptor to Defined Regions in the Human Growth Hormone and Placental Lactogen Genes.

Molecular Cloning of Bovine Viral Diarrhea Virus Sequences.
By A. Renard, F. Guilot, D. Schmetz, L. Dagenais, P. Pastorot, D. Dina and J.A. Martial

Hormone-Dependent Covalent Modification and Processing of Human Progesterone Receptors in the Nucleus.
By K. B. Horowitz, M.D. Francis, and L.L. Wet

An Highly Modular Cloning Vector for the Analysis of Eukaryotic Genes and Gene Regulatory Elements.
By D.S. Pfarr, G. Sathe, and M.E. Rief

Characterization of the Polypeptide Composition of Human Factor VIII.C and the Nucleotide Sequence and Expression of the Human Kidney cDNA.
By M.A. Truet et al.

An Ubiquitous Interspersed DNA Sequence Family in An Insect.
By J.Y. Bradfield, J. Locke, and G. R. Wyatt

Cloning and Isolation of Human Cytochrome P-450 cDNAs Homologous to Dioxin-Inducible Rabbit mRNAs Encoding P-450 4 and P-450 6.
By L.C. Quattrero, S.T. Okino, and U.R. Pendurthi

Structure of the Human Vasoactive Intestinal Polypeptide Gene.

Send subscription orders to: Mary Ann Liebert, Inc. publishers
157 East 86th Street, New York, N.Y. 10028 • (212) 289-2300

Note to authors: There are no page charges in DNA.
Volume 5, 1986, 6 issues: USA, $135; Overseas Air, $175
Volume 4, 1985, 6 issues: USA, $135; Overseas Air, $175
Volume 3, 1984, 6 issues: USA, $135; Overseas Air, $175
Volume 2, 1983, 4 issues: USA, $135; Overseas Air, $175
Volume 1, 1982, 4 issues: USA, $135; Overseas Air, $175
Forthcoming

Evolutionary Implication of Heterogeneity of the Nontranscribed Spacer Region of Ribosomal DNA Repeating Units in Various Subspecies of *Mus musculus*
Hitoshi Suzuki, Nobumoto Miyashita, Kazuo Moriwaki, Ryo Kominami, Masami Muramatsu, Takeharu Kaniehisa, François Bonhomme, Michael L. Petras, Ze-chang Yu and De-yuan Lu

Relative Contributions of Germline Gene Variation and Somatic Mutation to Immunoglobulin Diversity in the Mouse
Takashi Gojobori and Masatoshi Nei

Variation in the Number of α-Globin Loci in Sheep
A. Rando, L. Ramunno, and P. Masina

Review of *Isozymes: Current Topics in Biological and Medical Research, Volume 12*, Mario C. Rattazzi, John G. Scandalios, and Gregory S. Whitt, eds.
Thomas S. Whittam
A Comparison of the Small Ribosomal RNA Genes from the Mitochondrial DNA of the Great Apes and Humans: Sequence, Structure, Evolution, and Phylogenetic Implications
James E. Hixson and Wesley M. Brown

Conservation of a Long Open Reading Frame in Two Neurospora Mitochondrial Plasmids
Frank E. Nargang

The Heterochromatin of Grasshoppers from the Caledia captiva Species Complex. I. Sequence Evolution and Conservation in a Highly Repeated DNA Family
M. L. Arnold, R. Appels, and D. D. Shaw

Evolutionary History of the Hybridogenetic Hybrid Frog Rana esculenta as Deduced from mtDNA Analyses
Christina Spolsky and Thomas Uzzell

Polymorphism and Evolution of Influenza A Virus Genes
Naruya Saitou and Masatoshi Nei

Molecular Evolution of Bacteriophages: Evidence of Selection against the Recognition Sites of Host Restriction Enzymes
Paul M. Sharp

Charge Changes in Protein Evolution
Edward W. Peetz, Glenys Thomson, and Philip W. Hedrick

Book Review

Microorganisms as Model Systems for Studying Evolution (Monographs in Evolutionary Biology), edited by Robert P. Mortlock
Julian Adams

Books Received