A 16-kilobase region surrounding the transcription initiation site for ribosomal DNA and including the entire ribosomal DNA transcription unit has been characterized in man and compared in higher primates. Restriction analysis of ribosomal DNA from the pygmy chimpanzee (Pan paniscus), the common chimpanzee (Pan troglodytes), the gorilla (Gorilla gorilla), the orangutan (Pongo pygmaeus), the gibbon (Hylobates lar), and the rhesus monkey (Macaca mulatta) allows a primate phylogeny to be constructed based on ribosomal DNA structure. Individual variation and methylation are demonstrated in the ribosomal DNA repeats of all primates examined. Restriction analysis with HincII endonuclease suggests cleavage at sites containing methylated CpG and adds the SalI/HincII pair to those enzymes useful for studying DNA modification.

Introduction

Although primate relationships have been reconstructed from morphologic (Andrews and Cronin 1982), immunologic (Sarich and Wilson 1967), protein sequence (Wilson et al. 1977), and karyotypic (Miller 1977; Yunis and Prakash 1982) evidence, an evolutionary scheme that unifies these diverse phylogenetic data is not yet available. Recent studies of repeated DNA families (Gillespie 1977), mitochondrial DNA sequences (Brown et al. 1982), and the β globin gene cluster (Barrie et al. 1981) offer another approach to primate phylogeny, but novel molecular features pose new problems for evolutionary synthesis. The significance of rapid mitochondrial DNA mutation rates (Brown et al. 1979) or the sudden extinction of δ globin gene function (Barrie et al. 1981) will be clarified only by assembling structural data from a wide variety of primate genes.

The tandemly repeated genes for ribosomal RNA (rRNA) are attractive for molecular evolutionary studies because of their universal presence, high copy number, and functional similarity. Within each repeating unit are rapidly evolving
and highly conserved segments. In *Xenopus*, for example, rDNA nontranscribed spacer (NTS) sequences are widely divergent in sister species (Brown et al. 1972), while certain rRNA gene sequences are homologous to those of *E. coli* (Gourse and Gerbi 1980; Clark and Gerbi 1982). Comparisons between *Xenopus* and human rDNA again show great similarity between rRNA gene sequences but considerable differences between external transcribed spacer (ETS) segments (Wilson 1982). Ribosomal genes, then, should be useful for phylogenetic studies over long or short evolutionary distances and may recyle principles for repetitive DNA evolution that are applicable to diverse sequences and organisms. The detailed understanding of phylogenetic relationship and rDNA variation being assembled in *Drosophila* (Tartof 1979; Coen et al. 1982) may thus be helpful for understanding primate evolution through the medium of rDNA structure. In this report, we characterize the rDNA transcription initiation region of man and use cloned DNA segments from this region to compare its structure in seven primate species.

Human rDNA has been demonstrated by restriction analysis (Arnheim and Southern 1977; Schmickel et al. 1980) and electron microscopy (Wellauer and Dawid 1979) to consist of 44 kilobase (kb) repeating units with four EcoRI fragments labeled A–D (fig. 1). Molecular cloning of the EcoRI B fragment (Wilson et al. 1978), A fragment (Erickson et al. 1981), most of the C fragment (Higuchi et al. 1981; Meisfeld and Arnheim 1982), and a variable region of the D fragment (Erickson and Schmickel, in preparation) has provided probes for more extensive Southern analysis (Wilson et al. 1982) and DNA sequencing (Financsek et al. 1982; Meisfeld and Arnheim 1982). Some variable restriction sites within and among individuals have been defined, including an EcoRI site adjacent to the transcription initiation region (site E, in fig. 1), a HindIII site in the 28s rDNA gene (Arnheim et al. 1980), and length polymorphisms near the EcoRI site E (Wilson et al. 1982) and 3' to the 28s gene (Krystal and Arnheim 1978; Schmickel et al. 1980). The variable EcoRI and HindIII sites as well as the length polymorphism 3' to the 28s gene are distributed throughout the five nucleolar organizer regions in man, suggesting that exchanges between nonhomologous chromosomes may be one mechanism for maintaining rDNA homogeneity (Krystal et al. 1981). Certain of these polymorphisms have also been useful in preliminary comparisons of primate rDNA structure (Arnheim et al. 1980; Nelkin et al. 1980).

**Material and Methods**

**Primate DNA Isolation**

Human DNA was prepared from autopsy spleen or postnatal placenta specimens using the method of Marmur (1963). Nonhuman primate DNA was prepared from spleen, placenta, or blood samples provided by the Yerkes Primate Center. Lymphocytes were purified from primate blood specimens using the Ficoll-Hypaque method (Boyum 1975) prior to lysis and DNA purification (Marmur 1963). All specimens were dialyzed extensively in 10mM Tris-(hydroxyamino)methane hydrochloride, pH 7.5–0.1mM EDTA and stored over chloroform.

**Cloning and Preparation of Human rDNA**

Plasmid pBR 322 recombinant clones containing the human rDNA segments E,S,E,S,, or S,E were isolated as described previously (Wilson et al. 1982) and DNA prepared by the method of Bolivar and Backman (1979). Primary clones of the B,B BamHI segment of human rDNA were isolated from a human BamHI
Human rDNA Repeating Unit

\[ \text{BB BB}_1 \rightarrow \text{S}_1 \rightarrow \text{S}_2 \rightarrow \text{S}_3 \rightarrow \text{B}_4 \rightarrow \text{B}_5 \rightarrow \text{B}_6 \rightarrow \text{B}_7 \rightarrow \text{S}_4 \rightarrow \text{B}_8 \]

Detailed Map of \( E_2S_2 \) Segment

---

DNA Sequence of the \text{Sma I} B Fragment

\[
\begin{align*}
5'\text{GCGGAGGGA} & \text{TATCTTCTG} \text{CTCGAAGTCG} \text{GCAATTTGGG} \text{CCTACGAGTT} \\
\text{Sma I} & \\
\text{ATTGCTGAGA} & \text{CGGCTGACCT} \text{TGCGGA4GCT} \text{TGCGGAGAG} \text{GCTTGGGACCT} \\
\text{Ti} & \\
\text{GGGATGCGC} & \text{GCGGAGGTCG} \text{GCGCTACGGG} \text{TGGCGCCTA} \text{GCCGGCCGCC} \\
\text{GCTGCTGCT} & \text{GAGECGCTTG} \text{CAAGGGAGCC} \text{GGGGCTTGCC} \text{TTGTCCTTGC} \\
\text{GCCGTCGCA} & \text{GGCTCCGAC} \text{TCCGCTGCGG} \text{GCCC'F} \\
\end{align*}
\]

Fig. 1.—Characterization of the human rDNA transcription initiation region. A single repeating unit of human rDNA is shown in the upper panel with the four major \text{EcoRI} (E) fragments A-D. \text{BamHI} (B) and \text{Sall} (S) sites are also shown. The nontranscribed spacer (thin lines), transcribed spacer (thick lines), and gene regions (boxes) are indicated. The repeating unit is an estimated 44 kb of DNA repeated 100–300 times per haploid genome. The middle panel represents an enlargement of the 1,250-bp \( E_2S_2 \) segment showing relevant restriction sites and the \text{SmaI} fragments which have been cloned into bacteriophage M13 mp 9 for dideoxynucleotide sequencing. The DNA sequence of the \text{SmaI} B fragment containing the transcription initiation region (TI) as positioned 80 ± 1 bp upstream from a \text{BstE II} restriction site is shown in the lower panel. The underlined sequence is identical with that in the mouse. The DNA sequence supports the data of Meisfeld and Arnheim (1982) over those of Financsek et al. (1982) at positions 34, 53, 126–127, and 192 while supporting Financsek et al. (1982) at positions 5, 135, 175, and 186–188. Data from the latter laboratories agree on the deletion of C at positions 15 and 94, A at position 172, and the insertion of C at positions 84 and 224. The inset demonstrates the 80-bp \text{BstEII}-cleaved end-labeled fragment of \( E_2S_2 \) DNA (arrow) protected from S1 nuclease digestion by human 45S rRNA (lane 3). No protected fragment was demonstrated if bacterial tRNA was substituted for human 45S rRNA (lane 2). Lanes 1 and 4 show the end-labeled \text{SmaI} fragments A (290 bp), B (234 bp), C (170 bp), D (115 bp), E (60 bp), F (40 bp), G (23 bp), and H (20 bp) as size markers, and lane 5 demonstrates partial cleavage of the \text{SmaI} B rDNA fragment by \text{BstEII} to yield 130- and 105-bp fragments.

Library constructed in Charon 28 (Williams and Blattner 1980), using the procedure of Erickson et al. (1981). Approximately 50,000 recombinants were screened using as probe the plasmid clone \( E_2S_2 \), and the procedure of Benton and Davis (1977) to identify two B,B, recombinants. Bacteriophage DNA was prepared as described previously (Wilson et al. 1982).
Restriction Analysis of Primate DNA

Restriction endonuclease reactions (50 μl) containing 2–8 μg of primate DNA and 4–16 units of restriction enzyme (Bethesda Research) were incubated 4 h at 37°C before heating 5 min at 65°C. Electrophoresis was for 16 h at 5 V/cm on 0.8% agarose (Seakem) 15 cm (W) × 20 cm (L) horizontal gels using EcoRI and HindIII restriction fragments of lambda bacteriophage as size markers. Transfer to nitrocellulose (Schleicher and Schuell) was as described by Southern (1975) using 10 × SSC as the transfer medium. Hybridization in Seal-O-Meal bags (Sears Roebuck) was for 6 h at 65°C in 2 × SSC, 0.1% sodium dodecyl sulfate, 500 μg per ml sonicated E. coli DNA (Miles) followed by 15 h at 65°C with the same solution containing 1 μg of nick-translated DNA and 10% dextran sulfate. Nick-translations were as described (Rigby et al. 1977) using [α-32P]dATP (Amersham, 3,000 Ci/mmol) and 1 μg of bacteriophage or plasmid recombinant DNA. Specific activities ranged from 1 to 5 × 10^7 cpm per μg of plasmid DNA. After hybridization, nitrocellulose membranes were incubated 1 h at 65°C in 2 × SSC and 0.1% sodium dodecyl sulfate, then 3 h at 65°C in 3 × SSC before autoradiography at room temperature in cassettes containing a fluorescent intensifier screen and XR-5 Royal-X-O-Mat film (Kodak).

DNA Sequencing of the Human SmaI B Fragment

Random Smal subclones in the M13 bacteriophage vector mp 9 (Bethesda Research) were isolated from the E, S, segment and used as templates for dideoxynucleotide sequencing (Messing et al. 1981). Ligation reactions (50 μl) contained 1 μg of Smal-cleaved mp 9 DNA and 2 μg of Smal-cleaved E, S, DNA in ligation buffer (Wilson et al. 1978) with 1 unit of T4 DNA ligase (Bethesda Research). After overnight incubation at 4°C, 5-μl aliquots were incubated for 1 h at 4°C with 0.2 ml of shocked cells prepared from E. coli JM 103 (Bethesda Research). Transfectants were plated on yeast-tryptone indicator plates along with control reactions containing no ligase or no target DNA. Of 34 β-galactosidase negative “plaques” obtained, 27 exhibited nitrocellulose membrane hybridization to nick-translated pHrB-ES DNA after isolating single-stranded DNA by the polyethylene glycol method (Messing et al. 1981). Nine clones contained the Smal B fragment (fig. 1) as identified by BsrEII cleavage patterns of the replicative form. Other clones contained the Smal A (five), C (three), E (two), F (two), G (four), or H (one) fragments as identified by nucleotide sequencing, computer overlapping, and comparison to the human rDNA sequences of Financsek et al. (1982). One clone has not been characterized. Dideoxynucleotide sequencing of single-stranded DNA from the M13 Smal B fragment clones (three in each orientation) was performed using a 15-bp primer (P-L Biochemicals) as described (Messing et al. 1981). Sequencing reactions were applied to 40-cm (H) × 34-cm (W) × 0.4 mm 5% acrylamide-urea gels. Electrophoresis was for 3 h (1–120 bp) or 6 h (50–150 bp) at 1,500 V.

S1 Nuclease Protection Experiments

Hela cell suspension cultures in minimal media (GIBCO) with 10% newborn calf serum were harvested at 3–5 × 10^6 cells per ml. Nuclei and nucleoli were prepared and 45S rRNA was isolated from total cellular RNA by sedimentation on 15%–30% linear sucrose gradients as described (Weinberg et al. 1967).
According to the protocol of Berk and Sharp (1977), E,S, DNA (2-5 μg) was cleaved with HincII, BstEII, Xor, or SalI and 5' end-labeled with alkaline phosphatase (Sigma), polynucleotide kinase (Bethesda Research), and [γ-32P]-ATP (New England Nuclear). Purified 45S rRNA (1.2 μg) was hybridized with 100 ng (1 x 10^4 cpm) of 5' end-labeled, restricted E,S, DNA. Hybridizations were for 6 h at 60 C in 20 μl of 80% formamide, 0.04 M piperazone-N, N'-bis (2-ethanesulfonic acid), pH 6.4, 1 mM EDTA, 0.4 M NaCl after heating 10 min at 68 C. After diluting to 200 μl with S1 nuclease buffer (30 mM Na acetate, pH 4.6, 0.25 M NaCl, 1 mM ZnSO4), the hybrids were incubated for 1 h at 37 C with 100 units of S1 nuclease (Bethesda Research). After ethanol precipitation, electrophoresis was on 5% polyacrylamide gels with the end-labeled SmaI fragments A-H of E,S, (fig. 1) as molecular weight standards. The gels were dried under vacuum overnight (Hoefer apparatus); autoradiography was as described above.

Cross Hybridization of Cloned Human rDNA Segments

Since 10 μg of primate DNA should contain approximately 1 ng of rDNA sequences (Schmickel 1973), amounts of recombinant bacteriophage or plasmid DNA calculated to contain 1 and 5 ng of the cloned human rDNA segment were restricted with the appropriate enzymes (i.e., EcoRI and SalI for the clone E,S,) to separate insert and vector fragments after electrophoresis on 0.8% agarose gels. A panel of human rDNA clones (B,B,, E,S,, S2S3, S3E2, and E,E,) was included on each gel and hybridized to one of the respective nick-translated clones (five experiments) after Southern transfer as described above. Cross hybridization was striking among the vector fragments but nonexistent between the different rDNA inserts.

Synthesis of a strand-selective probe complementary to the SmaI B fragment (fig. 1) was accomplished using an M13 mp 9 recombinant clone as template that contained the SmaI B fragment in a 5'—3' orientation relative to the primer site (Hu and Messing 1982). Radioactive labeling was accomplished using the standard DNA sequencing reaction (Messing et al. 1981) with 10 μCi of [α-32P]dATP (2,000—3,000 Ci/mmol—Amersham) without dideoxynucleotides. After 90 min at 15 C, the reaction was stopped with EDTA. Single-stranded M13 recombinant DNA (0.1 μg) containing the rDNA SmaI fragments A, B, C, E, F, and G in the 5'—3' orientation relative to primer were bound to nitrocellulose in 1 M Na acetate using a dot-blot apparatus (Hybridot-Bethesda Research) and rinsed thoroughly with 2 x SSC. After hybridization with the SmaI B fragment probe under conditions described above, only the homologous SmaI B fragment DNA showed significant hybridization.

Parsimony Analysis of Primate rDNA Restriction Data

Computer-assisted parsimony analysis was kindly performed by Dr. W. Fitch using previously described methodology (Fitch 1977). The primate rDNA restriction maps were first reduced to nine characters (sites) whose states were coded by letters representing present, absent, or alternative restriction sites (see fig. 7). The 935 possible unrooted trees generated for these nine positions were examined in two ways. First, the H2/S2, H2/S1, and PvuII a/b/c alternatives (see fig. 7) were treated as equivalent sites requiring a minimum of 11 character state changes among the primates in the absence of homoplasy (parallel or back-mutations). Two most parsimonious trees were obtained by computer analysis, each requiring
16 changes. A second analysis treated $H_2/S_2$, $H_7/S_7$, and $PvuII$ a/b/c as separate sites requiring a minimum of 13 character state changes. Three most parsimonious trees were found, each involving 25 changes. The single most parsimonious tree common to both analyses is diagrammed in figure 7.

**Results**

**Characterization of the Human rDNA Transcription Initiation Region**

Prior to the comparison of primate rDNA structure, the human rDNA transcription initiation site was defined by molecular cloning, S1 nuclease mapping, and DNA sequencing. The 5.7-kb EcoRI B fragment of human rDNA (fig. 1) has been cloned in lambda bacteriophage vectors (Wilson et al. 1978, 1982) and EcoRI-SalI ($E_2S_2$), SalI-SalI ($S_2S_3$), and SalI-EcoRI ($S_3E_1$) segments subcloned in the plasmid pBR 322 (Wilson et al. 1982). The Smal fragments A–F within the $E_2S_2$ segment have been defined by partial restriction analysis (Wilson et al. 1982) and were end-labeled to provide size markers for S1 nuclease mapping experiments. Nuclease S1 cleavage of a hybrid between human 45S rRNA and the $E_2S_2$ segment restricted with BstEII yielded a protected fragment which localized the 5′-end of the rRNA 80 bp upstream of the BstEII site (see inset, fig. 1). Also protected from S1 digestion by purified 45S rRNA or nucleolar RNA were 700-bp and 350-bp fragments extending from the SalI or XorII sites, respectively, while no protected fragment was obtained when the $E_2S_2$ segment was end-labeled at the HincII site $H_4$ (data not shown). These experiments define a putative initiation site for rRNA transcription which agrees with results from other laboratories (Financsek et al. 1982; Meisfeld and Arnheim 1982). Similar experiments in the mouse (Grummt 1981; Miller and Sollner-Webb 1981) locate the rDNA transcription initiation site 5,200 bp from the EcoRI site $E_4$ in both mouse and man.

Templates for dideoxynucleotide sequencing of the human rDNA transcription initiation region were isolated by random subcloning of Smal fragments from the $E_2S_2$ segment into the M13 bacteriophage vector mp9 developed by Messing et al. (1981). The DNA sequence for the Smal B fragment is shown in figure 1 and defines a consensus sequence for the transcription initiation region based on results obtained by two independent techniques and three different laboratories (see legend to fig. 1). As pointed out previously (Financsek et al. 1982; Meisfeld and Arnheim 1982), the 15-bp sequence underlined in figure 1 is identical between mouse and man. It is also identical in the rat (Rothblum et al. 1982).

Before comparing primate rDNA transcription initiation regions, it was important to rule out possible cross hybridizability among the various human rDNA probes that would be used for Southern transfer experiments. Included in this analysis were two primary bacteriophage clones containing the human rDNA segment $B_2$, which were isolated in the BamHI vector Charon 28 (Williams and Blattner 1980). No cross hybridization among $B_2$, $E_2S_2$, $S_2S_3$, $S_3E_3$, or $E_3E_4$ human rDNA segments could be demonstrated using amounts and specific activities of cloned rDNA comparable to those in Southern transfer experiments with primate genomic DNA (data not shown). In addition, a probe constructed by biased labeling of the transcription initiation region (Hu and Messing 1982; see Material and Methods) did not hybridize to other Smal clones from the $E_2S_2$ segment or to the other rDNA segments mentioned above. These experiments suggest that duplication of promoter-like sequences which occurs in the *Xenopus*
rDNA NTS (Sollner-Webb and Reeder 1979) does not occur in the human rDNA 5'-NTS.

Comparative Restriction Analysis of Primate rDNA with *BamHI*

The basic topology of the transcription initiation region in several primate species is shown by *BamHI* restriction in figure 2A. The cloned segment E,S₂ is used as a probe to demonstrate 6.8-kb *BamHI* segments B₁B₂ (human, pygmy chimpanzee, and gorilla), 7.3-kb BₓB₂ segments (chimpanzee and gibbon), and variable BₓB₂ and/or 9.8-kb BₓB₂ segments in the rhesus monkey. These and subsequent restriction patterns represent analysis of 15 human, four Helu cell, seven chimpanzee, five pygmy chimpanzee, five gorilla, three gibbon, three orangutan, and eight rhesus DNA samples prepared as described in Material and Methods. Only the rhesus shows individual variation in the *BamHI* patterns, with each monkey having different proportions of the B₁B₂ and BₓB₂ fragments. Orangutan rDNA is unique in having a novel *BamHI* Bₓ site which divides the 6.8-kb B₁B₂

![Diagram of restriction analysis](image)

**Fig. 2.—Restriction analysis of primate rDNA with *BamHI*. A. Primate rDNA *BamHI* restriction patterns visualized by Southern transfer and hybridization to the cloned segment E,S₂. DNA from human (lane 1), pygmy chimpanzee (2), gorilla (3), chimpanzee (4), gibbon (5), rhesus (6, 7), and orangutan (8) was restricted with *BamHI* as described in Material and Methods. Molecular length standards for all experiments were *EcoRI* and *HindIII* restriction fragments of bacteriophage lambda. Positions of primate *BamHI* sites relative to human restriction sites are summarized in the lower panel and shown individually in fig. 7. B. Primate *BamHI* restriction patterns visualized by Southern transfer and hybridization to the cloned DNA segment S,S₂ (lanes 1–5) followed by S,E₁ (lanes 6–10). Human (lanes 1 and 6), gorilla (2 and 7), orangutan (3 and 8), gibbon (4 and 9), and rhesus (5 and 10) DNA restriction patterns are shown in comparison to lambda bacteriophage markers.
segment into 5.6-kb B,B, and 1.2-kb B,B, portions. The minor bands visible in lane 8 may represent "orphon" sequences (Childs et al. 1981) and have not been analyzed in detail.

The probes S,S, and S,ES, are used in figure 2B to demonstrate the 2.2-kb B,B, and 5.3-kb B,B, BamHI fragments which are conserved in all primates. This and similar experiments (Arnheim et al. 1980; see below) show that the variation among B,B,, B,B,, and B,B, segments is due to variation in the NTS and not in rRNA gene regions. As shown in figure 1, Higuchi et al. (1981) have defined a cluster of BamHI sites in the human NTS 5' to the B, site. This BamHI cluster is also evident in gorillas and other primates when partial BamHI restrictions are inadvertently obtained (lane 2, fig. 2B). The correspondence of primate sites B, and B, to these human BamHI sites, along with a lack of variation upon EcoRI restriction of the same region (Arnheim et al. 1980), suggests that the BamHI variation is a restriction rather than length polymorphism.

Comparative Restriction Analysis of Primate rDNA with SalI

Figure 3A shows a major 1.45-kb S,S, fragment and minor 14-kb S,S, and 3.7-kb S,S, fragments after SalI restriction of several human and Hela cell DNA samples using the cloned E,S, segment as hybridization probe. Since the S, and S, sites were present in all of 17 independent EcoRI B fragment clones of human rDNA (Wilson et al. 1982), the S,S, and S,S, fragments in figure 3A probably represent methylation of CpG nucleotides at the SalI S, or S, sites. The decreased amounts of the S,S, and S,S, SalI fragments in DNA from Hela cells support this hypothesis, since methylation is more variable in cell culture (Kunnath and Locker 1982). The SalI site S, shows a similar variation in man and other primates as demonstrated by 6.8-kb B,B, and 6.1-kb B,S, fragments visualized after BamHI/SalI restriction and hybridization to the cloned segment E,S, (data not shown). Methylation is again suggested as a cause for this variation since the SalI site S, is present in the two B,B, clones described above and in the eight clones isolated by Higuchi et al. (1981).

In figure 3B, the various primate species are compared for their SalI restriction patterns as demonstrated by hybridization to segment E,S, 2. The chimpanzee, pygmy chimpanzee, and rhesus monkey lack SalI S, sites, while all primates have S,S, and S,S, fragments. Additional S, or S, sites are seen in rhesus and orangutan, generating 3.15-kb S,S, and 2.7-kb S,S, fragments (lanes O and R, fig. 3B). A 0.55-kb S,S, SalI fragment is visible in orangutan and rhesus while a unique S,S, fragment is demonstrated in the orangutan. Different primate individuals of the same species give identical SalI restriction patterns except that the amounts of S,S,, S,S,, or S,S, fragments vary among orangutan and rhesus individuals. Sequential hybridization of the transfer in figure 3B with the cloned segment S,S, demonstrated additional 2.25-kb S,S, fragments in those species having an S, site, while additional hybridization with the segment S,ES, showed a 10.3-kb S,S, segment in all primates (data not shown). These results indicate a high degree of length and sequence conservation of the rDNA transcription unit within and among primate individuals.

The relationship of SalI and HincII restriction endonuclease recognition sequences diagrammed in figure 4 suggested that HincII restriction of primate rDNA might clarify the results of SalI restriction. Six restriction fragments are visualized by hybridization to the segment E,S, after partial restriction with HincII (lane 2,
Fig. 3.—Restriction analysis of primate rDNA with SalI and HincII. A. Human rDNA SalI restriction patterns demonstrated by hybridization to the probe E,S: Human spleen (lanes 1–3, human placenta (4–6), and Hela cell lines (Schmickel et al. 1980) S3 (7), Hep-2 (8), 222 (9). Detroit 96 (10) demonstrate the SalI rDNA fragments diagrammed in figs. 3 and 7. B. SalI restriction patterns of human (H), chimpanzee (C), pygmy chimpanzee (P), gorilla (G0), gibbon (G1), orangutan (O), and rhesus (R) rDNA after hybridization to E,S. C. HincII restriction of primate rDNA after hybridization with E,S. Lane 2 shows a partial HincII restriction of human rDNA with 3.15-kb H,S, 2.7-kb H,S, 1.8-kb H,H, 1.45-kb S,S, 0.9-kb H,S, and 0.55-kb S,H, fragments (see lower right). A complete human digest (lane 1) again demonstrates the 1.8-kb H,H, restriction fragment even after long periods of restriction and the expected H,S, and S,H, fragments. Partial restriction of chimpanzee rDNA (lane 3) yields the partial restriction fragments S,S, and H,H, as well as the S,H, fragment which remains after complete digestion. Complete digestion yields S,H, and S,H, fragments for pygmy chimpanzee (lane 4), S,S, S,S, and S,S, fragments for orangutan (lane 5), S,H, and S,S, (faint) for rhesus (lane 6).

fig. 3C). These fragments are labeled at the left margin and serve as reference markers for the other HincII patterns. Complete HincII restriction of human rDNA yields the minor fragment H,H, and the major fragments H,S, and S,H, (lane 1, fig. 3C). The human 3.15 H,S, fragment visualized in the partial HincII digest is absent from the complete digest (lane 1, fig. 3C), indicating the SalI site S, that
appeared absent or methylated in certain rDNA repeats (yielding an $S_1S_4$ fragment by SalI digestion) is cleaved by HincII. Given the previous analysis of the human EcoRI B clones, the simplest explanation is that HincII can cleave methylated SalI sites. Absence of the $S_1S_4$ fragment in pygmy chimpanzee rDNA (lane 4) and of the $S_1S_3$ or $S_2S_3$ fragments in orangutan (lane 5) or rhesus (lane 6) rDNA also
supports this conclusion. Of interest is the nonrandom \textit{HincII} cleavage of primate rDNA exemplified by the H$_1$H$_8$ fragment in the human rDNA complete digest (lane 1), the absent H$_8$S$_8$ fragment in the chimpanzee rDNA partial digest (lane 3), or the absent H$_8$H$_8$ fragment in the pygmy chimpanzee and rhesus restriction digests. This may reflect differences in rates of restriction or actual differences in restriction site combinations within individual rDNA repeating units (see Discussion). The positions of primate \textit{SalI} and \textit{HincII} sites are summarized in figures 3 and 7.

Results of two confirmatory experiments supporting the restriction maps in figures 3 and 7 are presented in figures 5 and 6. In figure 5, triplicate \textit{SalI} digests of human, chimpanzee, and rhesus DNA are hybridized separately to the probes E$_2$S$_2$ (A), S$_2$S$_3$ (B), and S$_3$E$_3$ (C). The rhesus rDNA fragment S$_2$S$_3$ (lane R, fig. 5C) is detected only by the probe E$_2$S$_2$, confirming the position of site S$_3$ as shown in figure 3. The lack of S$_2$S$_3$ fragments in all three primates using S$_2$S$_3$ as probe

![Image](image.png)

**Fig. 6.**—Restriction of primate rDNA with \textit{EcoRI} and \textit{SalI}. Primate rDNA samples are designated as in fig. 3B. Hybridization was with the DNA segment E$_2$S$_2$ followed by S$_2$S$_3$. The numbers refer to human rDNA restriction fragments 1-E$_2$E$_1$ (18 kb), 2-E$_2$E$_3$ (5.7 kb), 3-S$_2$E$_1$ (4.55 kb), 4-S$_2$S$_1$ (3.7 kb), 5-E$_3$S$_3$ (3.5 kb), 6-S$_2$S$_3$ (2.25 kb), 7-S$_3$S$_2$ (1.45 kb), and 8-E$_3$S$_2$ (1.25 kb). \textit{EcoRI}-\textit{SalI} restriction patterns of chimpanzee (C), gorilla (Go), gibbon (Gi), orangutan (O), and rhesus (R) are consistent with the maps in fig. 7.
demonstrates nonrandom methylation patterns within individual repeating units, as discussed below.

The relation of the variable human rDNA EcoRI site \( E_2 \) (Wilson et al. 1978; Arnheim et al. 1980) to modified or absent \( S_{al1} \) sites is examined by EcoRI/SalI restriction in figure 6. In lane 1, various human rDNA segments are demonstrated by sequential hybridization to the cloned segments \( E_S, S_r, E_S, S_r, \) and \( S_r, S_r, S_r, \). The occurrence of 1.45-kb \( S_r, S_r, \) and 1.25-kb \( E_r, S_r, \) segments (numbered 7 and 8 in fig. 6), 3.7-kb \( S_r, S_r, \) and 3.5-kb \( E_r, S_r, \) segments (numbered 4 and 5 in fig. 6), and the 4.55-kb \( S_r, E_r, \) segment (numbered 3 in fig. 6) suggests random combinations of variable \( E_r, S_r, \) and \( S_r, \) sites among individual human rDNA repeats, as would be expected if methylation were independent of the \( E_2 \) restriction site polymorphism. Similar patterns are shown for other primates in figure 6, and the results support the restriction maps diagrammed in figure 7.

Discussion

Molecular characterization of the human rDNA transcription initiation region (fig. 1) has provided a panel of rDNA segments for the analysis of primate restriction polymorphisms and has suggested that major rearrangement or duplication of promoter sequences has not occurred in the 75 Myr since divergence of man and mouse. Considerable length conservation of ETS and adjacent NTS regions can be demonstrated both within individuals and among seven primate species (figs. 2, 3). Even greater structural similarity of primate rDNA is observed within the rDNA transcriptional unit. Despite the overall picture of primate rDNA conservation, some specific mutations can be documented by comparing the hu-
man rDNA sequence with certain primate rDNA restriction sites. The *SalI* site GTCGAC labeled *S*, (figs. 3B, 7) in certain primates has become the *HincII* site *H*, in man (figs. 1, 3C, 7) with the sequence GTTGAC (Financsek et al. 1982; Meisfeld and Arnheim 1982). Similarly, the *SalI* site *S* in orangutan, gibbon, gorilla, and man (figs. 3B, 7) has become the *HincII* site *H* in rhesus, chimpanzee, and pygmy chimpanzee (figs. 3C, 7). As shown in figure 4, both changes represent transitions. Also, the *BamHI* site *B* in the orangutan (figs. 2, 7) has a homologous sequence GGGTCC in the human rDNA sequence (G. N. Wilson, unpublished; Financsek et al. 1982; Meisfeld and Arnheim 1982) and represents a transition. No homologue for the *SalI* *S* (figs. 3B, 7) is apparent in the human rDNA sequence.

The novel *BamHI* sites *B*<sub>x</sub>, *B*<sub>y</sub>, and *B*<sub>z</sub> (fig. 2) in nonhuman primates appear to represent restriction polymorphisms since *EcoRI* restriction shows no evidence of major length variation in primate 5' NTS regions (Arnheim et al. 1980). Also, the *B*<sub>x</sub> site seems to be present in most primates, as demonstrated by our own partial *BamHI* digests and by the cloning experiments of Higuchi et al. (1981). Yet some length heterogeneity in the 5'-NTS is undoubtedly present, since the primate *B*<sub>y</sub>*B*<sub>z</sub>, *B*<sub>x</sub>*B*<sub>z</sub>, and *B*<sub>y</sub>*B*<sub>z</sub> *BamHI* fragments shown in figure 2 appear as doublets after long periods of electrophoresis (not shown). Length variation on the order of 15–30 bp multiples surrounding the *EcoRI* site *F*<sub>2</sub> in man has been suggested by partial restriction analysis of *E*,*F*, clones (Wilson et al. 1982). Since these restriction and length variations have not been completely characterized, only the *B*<sub>x</sub> site is considered in the phylogeny.

A phylogeny was derived from the nine well-characterized primate restriction site differences diagrammed in figure 7. Using computer-assisted analysis of 935 possible unrooted trees (Fitch 1977), two most parsimonious trees involving 16 character state changes were defined counting the *H*/*S*, *H*/*S*, and *PvuII* a/b/c alternatives as equivalent sites, while three most parsimonious trees involving 25 character state changes were defined counting the alternatives as separate sites (see Material and Methods). The single most parsimonious tree common to both treatments is shown in figure 7. It agrees with primate phylogenies derived from comparisons of albumin (Sarich and Wilson 1967), mitochondrial DNA sequences (Ferris et al. 1981; Brown et al. 1982), or the β-globin cluster (Zimmer et al. 1980; Barrie et al. 1981). The rDNA phylogeny also supports the chromosomal data of Yunis and Prakash (1982) over those of Miller (1977) in suggesting that the chimpanzee is more closely related to man than is the gorilla.

It will be interesting to correlate the phylogeny in figure 7 with the dramatic changes in number and location of rDNA loci on primate chromosomes demonstrated by in situ hybridization and silver staining (Wilson 1982). Our preliminary results suggest that mechanisms for unequal crossing-over which are postulated to maintain rDNA homogeneity within individuals and species (Smith 1973) may lead to sudden changes in rDNA location and dosage without affecting molecular structure. Further conclusions regarding the tempo and mode of primate rDNA evolution will, of course, require restriction analysis of the entire repeating unit and DNA sequence comparisons of selected regions so that bias can be minimized (Adams and Rothman 1982). Also important, as will now be discussed, is better knowledge of the extent and mode of rDNA variation within primate individuals and species.
A notable feature of the phylogeny in figure 7 which may relate to individual variation is the transition from the \textit{SalI} site \textit{S}, to the \textit{HincII} site \textit{H}, that occurs in three separate primate branches. These changes suggest parallel evolution which may be related to variable modification of this site as demonstrated in figure 3. Approximately 10\%-20\% of human rDNA repeating units from placenta or spleen are not cleaved by \textit{SalI} at the \textit{S}, site, as compared with 2\%-5\% of HeLa cell repeating units (fig. 3A). In all cases, the \textit{SalI} \textit{S}, site is cleaved by \textit{HincII} (fig. 3C), which suggests two possibilities for the \textit{SalI} variation. Certain human rDNA repeats may have developed the \textit{HincII} site \textit{H}, which is present in other primates (fig. 7). This is unlikely, however, since 17 cloned \textit{E}$_{1}$\textit{E}, segments from two individuals all had the \textit{S}, site (Wilson et al. 1982). Also, there is variable cleavage by \textit{SalI} at sites \textit{S}, \textit{S}, (all primates—see Results), \textit{S}, or \textit{S}, (orangutan, rhesus); complete \textit{HincII} restriction occurs at each of these sites. An explanation more consistent with these results is that the \textit{SalI} variations represent methylation of \textit{CpG} which is cleaved by \textit{HincII}.

Although direct evidence examining the ability of \textit{HincII} to cleave methylated \textit{CpG} has not been published to our knowledge, this explanation seems reasonable, since the sequence GTTGAC with a similarly positioned thymine methyl group is recognized by \textit{HincII} (fig. 7). Assuming that methylation is responsible for the \textit{SalI} variation, then the results in figures 3 and 5 demonstrate nonrandom methylation within the \textit{S},\textit{S}, region of individual rDNA repeats. Figure 2A demonstrates that site \textit{S}, alone or both \textit{S}, and \textit{S}, may be methylated in individual human rDNA repeats. Methylation of site \textit{S}, alone, however, as would have been indicated by the presence of a \textit{S},\textit{S}, fragment, apparently does not occur (fig. 5). Similar selectivity of methylation within individual rDNA repeats of the chimpanzee and rhesus monkey is demonstrated in figure 5. It should be noted that the apparent absence of \textit{SalI} site \textit{S}, in chimpanzee or pygmy chimpanzee might represent 100\% methylation, although this seems unlikely. These nonrandom patterns of rDNA methylation will be interesting to examine in rapidly proliferative versus slowly growing tissues and in individuals with amplified and highly methylated nucleolar organizer regions (Tantravahi et al. 1981).

Another type of heterogeneity among individual rDNA repeating units, which in at least one case is quite distinct from rDNA methylation (fig. 6), is revealed by restriction analysis of primate rDNA. The \textit{EcoRI} sites \textit{E}, and \textit{E}, vary among individual humans, pygmy chimpanzees, or chimpanzees; the \textit{BamHI} sites \textit{B}, and \textit{B}, vary within rhesus individuals; and a \textit{HindII} site in the 28S rRNA gene varies among all primates (Arnheim et al. 1980). The subset of predicted \textit{HincII} cleavage fragments shown in figure 3C may also represent nonrandom restriction site combinations, although different cleavage rates at certain sites is also possible. These preliminary data suggest that the rDNA of a primate individual comprises a large number of structurally different repeating units which must be augmented or corrected in germ cells to explain the concerted evolution (Arnheim et al. 1980) demonstrated in figure 7. We propose the term “repetype” to describe alternate repeating unit structures within an individual.

If one considers the variable \textit{EcoRI} sites \textit{E}, and \textit{E}, (fig. 6), the \textit{HindII} site in the 28S rRNA gene (Arnheim et al. 1980), and the 4–5 length polymorphisms caused by a variable insertion in the 3'–NTS (Schmickel et al. 1980), then as many as 80 repetypes may be present in the chimpanzee, an organism having an estimated 488 rDNA repeating units (Wilson et al. 1982). Since relatively few detailed
restriction studies of primate rDNA have been performed, it is possible that every rDNA repeating unit has different combinations of restriction sites reflecting primarily variation in spacer regions. Further enumeration of rDNA repotypes in somatic and germ-line tissues will be required before the role of intra- (Smith 1973) or interchromosomal (Arnheim et al. 1980) crossover in maintaining rDNA homogeneity can be quantitatively understood. Restriction analysis under conditions where large DNA fragments can be resolved and the cloning of multiple "identical" rDNA segments from a single tissue source (Erickson et al. 1981; Wilson et al. 1982) are two approaches which can be followed. The multiplicity of rDNA repotypes in different cell types of an individual, their inheritance, and their distribution among populations will be important to define before a thorough understanding of rDNA evolution can be achieved.

Acknowledgments

We thank Dr. Walter Fitch for helpful criticism and for performing the parsimony analysis. Grant support from National Institutes of Health (PO1 HDGM 13506) and the National Foundation–March of Dimes (5-169 to G.N.W.) is also gratefully acknowledged.

LITERATURE CITED


WALTER M. FITCH, reviewing editor

Received July 29, 1983; revision received November 14, 1983.
Silencing of Duplicate Genes: A Null Allele Polymorphism for Lactate Dehydrogenase in Brown Trout (*Salmo trutta*)

Fred W. Allendorf
University of Montana

Gunnar Ståhl and Nils Ryman
University of Stockholm

A previously described isozyme polymorphism at one of two skeletal muscle *LdhA* loci in brown trout is due to a null allele, *Ldh1(n)*, producing no detectable catalytic activity. Homozygotes for this allele have approximately only 56% of the LDH activity in skeletal muscle relative to homozygotes for the active allele. The remaining activity results from enzyme subunits produced by other LDH loci. The *Ldh1(n)* allele is common and widespread throughout brown trout populations in Sweden and is also found in populations from Ireland. The persistence of duplicate gene expression for the *LdhA* loci in almost all salmonid species is best explained by natural selection against individuals containing null alleles. However, there is no indication of natural selection against brown trout with the *Ldh1(n)* allele. We suggest that the selection against individuals containing null alleles that is apparently responsible for the persistence of duplicate *LdhA* loci in salmonids occurs only under certain environmental conditions.

Introduction

The incorporation of duplicate genes into the genomes of vertebrates has played an important evolutionary role (Ohno 1970; Fisher et al. 1980). Recent studies with several groups of polyploid fishes have examined the structural and regulatory divergence of many pairs of duplicated loci (reviewed in Li [1980]). A common finding is that approximately half of the duplicate genes no longer produce a detectable enzyme product. The loss of duplicate gene expression is thus apparently a frequent occurrence.

Haldane (1933) first suggested that one of two duplicate loci may become nonfunctional through the fixation of mutations at one locus, while the other locus continued to perform the original function. This process was first mathematically treated deterministically by Fisher (1935) and stochastically by Nei and Roychoudhury (1973). The studies of duplicate genes in polyploid fishes have stimulated a series of recent papers theoretically examining the expected rate of loss,

1. Key words: polymorphism, salmonids, lactate dehydrogenase, gene duplication, null alleles.

Address for correspondence and reprints: Fred W. Allendorf, Department of Zoology, University of Montana, Missoula, Montana 59812.
by any means, of duplicate gene expression by the fixation of alleles producing no detectable enzyme, that is, null alleles. The different models described in those papers have been reviewed by Li (1980). These papers have raised several important questions relating to the process resulting in the observed gene silencing. An important consideration in discriminating among possible models is the frequency of null alleles at loci that have retained duplicate gene expression (Li 1980).

We have described an electrophoretic polymorphism for a skeletal muscle lactate dehydrogenase (LDH, EC 1.1.1.27) locus (Ldh1) in brown trout (Salmo trutta) that is due to an allelic product that either comigrates with the products of the other muscle LDH locus (Ldh2) or has no enzyme activity (Allendorf et al. 1976). This paper shows that this polymorphism is caused by a null allele that is widespread and is present in high frequencies in some Swedish populations of brown trout.

Material and Methods

The samples were collected for various purposes in an ongoing study of genetic variability in natural populations and hatchery stocks of brown trout in Sweden. The geographic locations of the sample sites are shown on the map in figure 1. Most lakes are in alpine areas in central and northern Sweden. The majority of fish were caught with gill nets of varying mesh sizes. A section of skeletal muscle was taken from behind the head of every specimen and brought to the laboratory for freezing (−60 C) until electrophoresis was performed.

![Map of Sweden with locations marked](image-url)

Fig. 1.—Location of population samples containing (open circles) and lacking (crosses) Ldh1 null allele homozygotes in Swedish brown trout. Data for the 14 individual Swedish circles and the 40 collective crosses are shown in table 2.
Tissue extracts were prepared and horizontal starch-gel electrophoresis was done as described by Utter et al. (1974). Two different buffer systems were used: A, described by Ridgway et al. (1970), pH 8.5; and B, described by Clayton and Tretiak (1972), pH 6.1. Staining procedures followed Allendorf et al. (1977).

The serial dilution method of Klebe (1975) was used to estimate the relative amounts of LDH enzyme activity for different Ldh1 phenotypes at 37 C. The amount of LDH activity in muscle tissues was estimated by following the rate of nicotinamide adenine dinucleotide (NAD) reduction using the procedures of Neilands (1955) with crude skeletal muscle homogenates diluted 1:300 in 0.1 M phosphate buffer, pH 7.0, at 25 C. Three separate skeletal muscle samples were taken and analyzed from each individual. Individuals are compared on the amount of LDH activity per unit weight of skeletal muscle tissue.

Results

Tissue Specific Expression

Five LDH loci have been described in trout, salmon, and char (Wright et al. 1975; Bailey et al. 1976). Ldh5, producing the most anodally migrating enzyme, is expressed only in the eye of brown trout (fig. 2). Ldh4 is expressed in all tissues examined. Ldh3 is expressed in most tissues, with the exception of liver, and encodes the predominant form of LDH in heart tissue. The Ldh1 and Ldh2 loci are expressed only in skeletal muscle. Ldh1 and Ldh2 are paralogous genes that are orthologous to the LdhA locus of other vertebrates and were presumably duplicated in the ancestral salmonid (Bailey and Wilson 1968; Markert et al. 1975). Similarly, the Ldh3 and Ldh4 loci are paralogous duplicates arising from the orthologous LdhB locus found in many groups of fish (Wright et al. 1975; Markert et al. 1975). Only one of the duplicates of the LdhC locus produced by tetraploidy can still be detected (Ldh5).

![Image of LDH isozymes distribution in tissues of brown trout.](attachment:image.png)

**Fig. 2.**—Distribution of LDH isozymes in tissues of brown trout. The location of the homotetramers produced by the common allele at each of the five loci is shown. W = white skeletal muscle, H = heart, L = liver, E = eye, B = brain, S = stomach, K = kidney.
Variation at *Ldh1*

We originally described a variant phenotype consisting of only a single isozyme for the two *LdhA* loci (*Ldh1* and *Ldh2*) (Allendorf et al. 1976). This phenotype was assumed to result from a variant *Ldh1* allele that either produced a homotetrameric isozyme that comigrated with the products of the *Ldh2* locus or produced no detectable enzyme product. The absence of the intermediate phenotype expected of a heterozygote in this sample has been shown to result from the presence of two subpopulations that are apparently fixed for the two different alleles (Ryman et al. 1979). We have since detected apparent heterozygous phenotypes in other populations (fig. 3).

There are two basic ways to distinguish between the two possible genetic models for the observed phenotypes. First, one could test to see if the single band of muscle LDH activity in the alternative homozygotes actually consisted of different isozymes produced by both *Ldh1* and *Ldh2*. We have not been able to separate this single band into multiple isozymes using several other buffer systems in addition to the two described in this paper. Any more sensitive biochemical procedures (e.g., comparative amino acid composition) are not likely to provide conclusive evidence for the presence of LDH1 enzyme because the *Ldh1* and *Ldh2* products are very similar (Lim et al. 1975) and the analysis would have to be done on the putative mixture of these products present in the single electrophoretic band.

We have therefore pursued the second approach, determining if there are differences in the amount of LDHA activity in the muscle of fish having different

![Image](image-url)  

**Fig. 3.**—The three phenotypes detected at the *Ldh1* locus: a = 100/100, b = 100/n, c = n/n
Ldhl phenotypes. We first did this by comparing the amount of LDHA activity in the two homozygous phenotypes using the visual end-point method of Klebe (1975). This procedure has been found to produce results with salmonid LDHA isozymes essentially identical to those obtained by densitometry of the bands on a starch gel under conditions in which the formazan deposition is linearly proportional to the enzyme units of LDH present (Lim and Bailey 1977). This was done using electrophoretic buffer system B because all five LDHA isozymes comigrate under those conditions. Comparison of 20 common homozygotes and five variant homozygotes showed that there is approximately twice the amount of LDHA activity per unit weight of muscle tissue in fish that are common homozygotes. This finding supports the model of the variant allele producing no detectable enzyme activity.

To estimate more accurately the relative amount of LDHA activity, we assayed the relative rate of reduction of NAD by spectrophotometric measurement in three separate crude muscle extracts from each of five fish of each phenotype (table 1). The heterozygotes and null homozygotes had an average of 88.7% and 55.8% of the average specific activity of the common homozygotes. A variance analysis of these data estimates 3.8% of the variation to be due to differences between samples within individual fish, 5.0% between fish with the same genotype ($F[12,29] = 4.91, P < .001$), and 91.2% between genotypes ($F[2,12] = 73.43; P < .001$). Therefore, we conclude that this polymorphism is caused by a null allele, Ldhl(n), because of the nearly 45% reduction in total muscle LDH activity in fish having the null homozygous genotype. Remaining activity results from enzyme subunits by other LDH loci.

A variety of potential genetic mechanisms for producing a null allele are possible: suppression of transcription or translation, or the production of an inactive enzyme via a nucleotide substitution, premature chain termination, or a deletion. Currently we cannot discriminate among these possibilities. However,

**Table 1**

Relative LDH Activities in Skeletal Muscle of Brown Trout Showing Three Electrophoretic Phenotypes at the Ldhl locus

<table>
<thead>
<tr>
<th>FISH NO.</th>
<th>Ldhl PHENOTYPE</th>
<th>RELATIVE ACTIVITY IN SAMPLE</th>
<th>MEAN ACTIVITY</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>III</td>
</tr>
<tr>
<td>1</td>
<td>100/100</td>
<td>.97</td>
<td>1.05</td>
</tr>
<tr>
<td>2</td>
<td>100/100</td>
<td>.98</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>100/100</td>
<td>.95</td>
<td>.87</td>
</tr>
<tr>
<td>4</td>
<td>100/100</td>
<td>1.02</td>
<td>1.11</td>
</tr>
<tr>
<td>5</td>
<td>100/100</td>
<td>1.03</td>
<td>1.14</td>
</tr>
<tr>
<td>6</td>
<td>100/n</td>
<td>.86</td>
<td>.87</td>
</tr>
<tr>
<td>7</td>
<td>100/n</td>
<td>.84</td>
<td>.86</td>
</tr>
<tr>
<td>8</td>
<td>100/n</td>
<td>.83</td>
<td>.94</td>
</tr>
<tr>
<td>9</td>
<td>100/n</td>
<td>.90</td>
<td>.92</td>
</tr>
<tr>
<td>10</td>
<td>100/n</td>
<td>.91</td>
<td>.91</td>
</tr>
<tr>
<td>11</td>
<td>n/n</td>
<td>.55</td>
<td>.64</td>
</tr>
<tr>
<td>12</td>
<td>n/n</td>
<td>.48</td>
<td>.50</td>
</tr>
<tr>
<td>13</td>
<td>n/n</td>
<td>.61</td>
<td>.62</td>
</tr>
<tr>
<td>14</td>
<td>n/n</td>
<td>.54</td>
<td>.44</td>
</tr>
<tr>
<td>15</td>
<td>n/n</td>
<td>.61</td>
<td>.55</td>
</tr>
</tbody>
</table>

**Note.—** A value of 1.00 is the mean amount of activity in the 100/100 homozygotes per gram of tissue.
intensive screening for null mutations of alcohol dehydrogenase and xanthine dehydrogenase in *Drosophila melanogaster* has detected many null mutants, almost all of which have been found to produce inactive proteins (Gelbart et al. 1976; Schwartz and Sofer 1976).

Geographic Distribution

The *Ldh1* null allele is widespread in brown trout populations throughout Sweden (table 2 and fig. 1). Because of the difficulty in distinguishing heterozygotes, we have treated this system as a simple recessive when estimating allelic frequencies from natural populations. Table 2 includes all population samples in which more than 30 individuals were examined for *Ldh1* variation. There is no apparent pattern in the geographical distribution of allele frequencies of these 54 population samples.

Discussion

Brown trout are polymorphic for the number of loci coding for LDH. At least one population, Lake Bumnersjoarna (deme II of Ryman et al. [1979]), has apparently become fixed for the null allele at *Ldh1* so that these fish have only four loci producing LDH. Populations throughout Sweden are polymorphic for the null allele at intermediate frequencies.

It is difficult to get a clear picture of the distribution and frequency of the null allele because it is treated as a recessive allele. Fourteen of the 54 samples contained at least one *Ldh1(n/n)* homozygote. The probability of not detecting a null homozygote with an allele frequency of *q* with a sample size of *N* is \((1 - q^2)^N\). Thus, there is a 61% chance of not detecting a null allele with an allele frequency of 0.10 with a sample size of 50 individuals. These results are thus

<p>| Table 2 |
| Frequencies of the <em>Ldh1</em> Null Allele in PopulationSamples of Brown Trout from Sweden |</p>
<table>
<thead>
<tr>
<th>Sample Site</th>
<th>Genotypes</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ldh1 (other)</td>
<td>Ldh1 (n/n)</td>
<td>Ldh1(n/n)*</td>
</tr>
<tr>
<td>1. Bumnersjoarna, deme II</td>
<td>0</td>
<td>155</td>
</tr>
<tr>
<td>2. Hetenjaure</td>
<td>65</td>
<td>43</td>
</tr>
<tr>
<td>3. Österjärn</td>
<td>85</td>
<td>15</td>
</tr>
<tr>
<td>4. Östra Vattensjön</td>
<td>42</td>
<td>6</td>
</tr>
<tr>
<td>5. Storan</td>
<td>30</td>
<td>4</td>
</tr>
<tr>
<td>6. Fällevatnet</td>
<td>66</td>
<td>8</td>
</tr>
<tr>
<td>7. Kaitum älv</td>
<td>48</td>
<td>3</td>
</tr>
<tr>
<td>8. Bolagen</td>
<td>34</td>
<td>2</td>
</tr>
<tr>
<td>9. Grubbvatnet</td>
<td>46</td>
<td>2</td>
</tr>
<tr>
<td>10. Vålåsjön</td>
<td>129</td>
<td>4</td>
</tr>
<tr>
<td>11. Munsvatnet</td>
<td>47</td>
<td>1</td>
</tr>
<tr>
<td>12. Stensjöflyarna</td>
<td>122</td>
<td>2</td>
</tr>
<tr>
<td>13. Stensjön</td>
<td>68</td>
<td>1</td>
</tr>
<tr>
<td>14. Lågsjön</td>
<td>159</td>
<td>2</td>
</tr>
<tr>
<td>40 samples</td>
<td>730</td>
<td>0</td>
</tr>
</tbody>
</table>

*NOTE.—* Only samples of 30 or more are included.

* This is the square root of the ratio of *Ldh1(n/n)* to the sum of both genotypes.
consistent with the null allele being widespread in Swedish populations of brown trout. An allele producing a similar one-banded phenotype has also been detected in a population of brown trout in Ireland (Taggart et al. 1981). Such an allele is thus apparently widespread in brown trout throughout northern Europe.

Salmonid \textit{LdhA} Loci

Lim and Bailey (1977) have described some biochemical properties of the isozymes produced by the two \textit{LdhA} loci of chinook salmon (\textit{Oncorhynchus tshawytscha}) and brown trout. They found no significant differences between the \textit{Ldh1} and \textit{Ldh2} homotetrameric isozymes of brown trout in several catalytic properties (Michaelis constants, substrate [pyruvate] optima, and resistance to product inhibition by L. [+\textsuperscript{+} - lactate]). They concluded that the products of these two loci are probably catalytically equivalent enzyme subunits in vivo. They did find, however, that the subunits produced by the \textit{Ldh1} and 2 loci of chinook salmon differ by some 12–25 amino acid residues per subunit. This is similar to the amount of divergence they had previously reported between the two \textit{LdhB} loci of chinook salmon (eight to 25 residues per subunit); this is equivalent to the amount of divergence expected after approximately 100 million years of divergence, if one assumes equal and constant rates of evolution (Lim et al. 1975).

Most salmonids apparently have retained both \textit{LdhA} loci. The arctic grayling (\textit{Thymallus arcticus}) and pygmy whitefish (\textit{Prosopium coulteri}) are the two possible exceptions. Only a single LDH1A isozyme has been detected in these species (unpublished data). Electrophoretic variation for LDHA in grayling indicates that only a single locus codes for this enzyme (unpublished data). Lynch and Vyse (1979), however, claim to have detected evidence of two \textit{LdhA} loci in the Arctic grayling. Klar and Stalnaker (1979) have reported an apparent null allele at the \textit{Ldh2} locus of cutthroat trout (\textit{Salmo clarki}).

Lim and Bailey (1977) have suggested that one of the \textit{LdhA} loci in Pacific salmon (\textit{Oncorhynchus} spp.) may be in the process of gradually being lost by the accumulation of regulatory mutations that reduce the synthesis of the products of one of the loci. However, such mutations resulting in a small reduction in gene expression should occur and be incorporated at both loci. This process would result in the retention of both loci with reduced expression. We would suggest, therefore, that the accumulation of such regulatory mutants reducing expression may in fact be responsible for the retention of the duplicate \textit{LdhA} loci in salmonids.

Frequency of Null Alleles

Enzymatically inactive alleles are rare in populations of diploids. Null alleles have been found to occur at a mean frequency of 0.0024 at 20 autosomal loci in two populations of \textit{Drosophila melanogaster} (Voelker et al. 1980; Langley et al. 1981). Mean frequencies of null alleles similar to this have been found in populations of ponderosa and red pine (0.0031 at 29 loci in \textit{Pinus ponderosa} and 0.0028 at 27 loci in \textit{P. resinosa}; Allendorf et al. [1983a]). The low frequency of null alleles is apparently best explained by some reduction in fitness of heterozygotes for null alleles (Langley et al. 1981). Loci that produce enzymes whose in vivo activities are not well known (e.g., esterases) are an exception to the dearth of null alleles in diploids (Burns and Johnson 1967).

As expected, null alleles are more common at the many duplicated loci in salmonid fish (Stoneking et al. 1981). Engel et al. (1973) have described a null
allele polymorphism at one of two apparently functionally equivalent LDH loci in the polyploid carp (*Cyprinus carpio*). Null alleles are apparently quite rare, however, at duplicated enzyme loci in the tetraploid derived catostomids (Ferris and Whitt 1979). This difference between the catostomids and salmonids suggests that the process of gene silencing continues in salmonids but not catostomids. This conclusion is in agreement with observations of duplicate loci in salmonids that still have not completed the process of diploidization (Wright et al. 1983); this is in contrast to the catostomids in which all duplicate loci show structural and/or regulatory divergence (Ferris and Whitt 1979). Null alleles may also be rare in catostomids because of smaller population sizes (Li 1980). Li (1980) has shown that “during the course of gene silencing the population is usually monomorphic for the normal allele if the population is small, but is polymorphic the majority of the time if the population is large.”

**Models for Loss of Duplicate Gene Expression**

The *LdhA* loci are unusual in comparison with other duplicated loci in salmonids. The loss of expression of one of the duplicate genes has been a common occurrence (Allendorf et al. 1975; Allendorf 1978; May et al. 1980). Those pairs where both loci retain expression usually show one of two characteristics. Either the loci have evolved different tissue-specific expression or the loci have not yet completely evolved disomic inheritance (Wright et al. 1981; May et al. 1982). In the first case the loci are no longer “functional” duplicates because some tissues have only one or the other expressed. Thus, the loss of activity of one locus cannot be compensated for by the other locus in those tissues in which they are not both expressed. In the second case, both loci are expected to have similar frequencies of alleles because they cannot diverge until they have become genetically distinct via disomic inheritance. Therefore, a null allele cannot be common at one locus without being present at the other locus as well. Thus, selection against the null allele would remain effective at both loci.

The *LdhA* loci fit neither of these characteristics. There is no evidence of any tissue-specific differences between the two loci. The estimated time of divergence of the two loci based on amino acid differences (Lim et al. 1975) is evidence that these loci have been diverging for a very long time. The retention of both loci is therefore somewhat of a puzzle. We next consider those models for loss of duplicate gene expression that have been presented.

Allendorf (1979) has considered a model in which there is selection for a reduced amount of enzyme activity at duplicate loci. As expected, loss of duplicate gene expression occurs quite rapidly with this model. This model, therefore, does not apply to salmonid *LdhA* in which duplicate gene expression has been retained in almost all species.

Many authors have considered the rate of loss of duplicate gene expression when all genotypes have equal fitness, except for a reduction in the fitness of the double homozygotes for a null allele (reviewed in Li [1980]). Even with this model, it is highly unlikely that duplicate gene expression would be retained after millions of years of divergence of the two loci. For example, with a null mutation rate of $10^{-8}$ and a population size of 10,000, the expected median number of generations until loss of duplicate gene expression is 120,000 generations (Li 1980). Li (1980) has argued that the slow rate of gene silencing in salmonids can be at least partially explained by recent establishment of disomic segregation or regulatory divergence...
of the two genes. As discussed previously, neither of the conditions apparently applies to salmonid *LdhA* genes.

We are therefore left with a third model of gene silencing in which all genotypes containing the null allele are selected against (Takahata and Maruyama 1979). These authors believe that this pattern of selection may be responsible for the slow rate of gene silencing in salmonids. This pattern of selection is apparently the best explanation for the retention of both *LdhA* loci in almost all salmonid species.

Can this pattern of selection be reconciled with the observed distribution of a null allele at *Ldh1* in brown trout? The high frequency and widespread occurrence of the *Ldh1(n)* allele would indicate that this allele is currently not being strongly selected against. However, the effects on fitness of a null allele will vary with species and environmental conditions. For example, reduced amounts of muscle LDH activity may have no effects on trout living in a hatchery or in a lake. Such reductions, however, may be significant in trout living or making spawning migrations through fast-flowing streams and rivers. These ideas could be tested directly by comparing the endurance of fish with and without LDH1 activity. Studies with the rainbow trout have shown that the absence of the activity of enzymes produced by individual loci can have important phenotypic effects (Allendorf et al. 1983b; Leary et al. 1983).

**Acknowledgments**

We thank Prof. Bryan Clarke for the use of laboratory facilities and Paul Barrow for his excellent technical help in the measurement of LDH activities. Some of this work was done while the senior author was supported by NATO and European Molecular Biological Organisation fellowships. During preparation of this manuscript, the senior author was supported by National Science Foundation grants DEB-8004681 and ISP-8011449. Part of this investigation was supported by grants from the Kungliga Fysiografiska Sallskapet in Lund and the Swedish Natural Science Research Council (B 3746-004 and -006).

**LITERATURE CITED**


Masatoshi Nei. reviewing editor

Received July 27, 1983; revision received September 14, 1983.
According to 5-Myr-old fossil evidence, ground squirrels within the genus *Spermophilus* had diverged into subgenera *Spermophilus* and *Otospermophilus* by late Miocene times. Radiometric dating has also provided a precise time for the sudden onset of a geological event, occurring 0.725 Myr ago, that initiated the complete and permanent reproductive isolation of two subspecies within the subgenus *Otospermophilus*. Since these two subspecies (S. *beecheyi* and S. *b. douglasii*) readily hybridize with each other under laboratory conditions, allopatric subspeciation is unlikely to have occurred prior to 0.725 Myr ago. We employed Nei’s model for estimating genetic distance in units which are linear in time, calibrated on the 0.725-Myr-ago date for initiation of *S. b.* subspeciation, to test its ability to generate a time scale for subgeneric divergence in keeping with the minimum estimate provided by the fossil record. This represents the most valid test to date of the utility of Nei’s model for estimating genetic distance in units which are linear in time. Nei’s model was found to underestimate this minimum time by 1 Myr, but it approximated this date after correcting values of D for variation in rates of evolution among loci.

**Introduction**

Kimura (1968) hypothesized that the principal cause of molecular evolution is the random fixation of selectively neutral mutations at rates that are constant over time. Various methods have been employed to estimate genetic distance between taxa in units of gene substitution as a means of determining the time of divergence (Langley and Fitch 1974; Nei and Roychoudhury 1974; Sarich 1977). In such cases, the mathematical model for estimating genetic distance must be one which insures linearity with respect to time. Estimates made by these methods are subject to at least four sources of error: (1) variation among loci in neutrality of or functional constraints on gene substitutions, (2) differences in laboratory methods for detecting variation (King and Wilson 1975; Bruce and Ayala 1979), (3) variation in the rates of neutral substitutions among those loci (King and Jukes 1969), and (4) variation in the rate of molecular evolution among groups of taxa.
1969; Nei and Chakraborty 1973; Langley and Fitch 1974), and (4) variation in substitution rates among the taxa selected for study (Wyles and Gorman 1980). Linearity of the distance measure in time might be further insured by comparing taxa with similar generation lengths (see Korey 1981), although an influence of differences in generation length among the taxa compared upon annual rates of gene substitution has been disputed (see, e.g., Wilson et al. 1977). Finally, the molecular clock has usually been calibrated indirectly using radiometrically dated fossils which have both already diverged and experienced a gradual cessation of gene flow (Cronin et al. 1980; Cronin and Meikle 1982), using model parameters whose values are estimated with substantial error (Fitch and Langley 1976) or employing imprecisely dated or lengthy geological events (Radinsky 1978), such as the emergence of the Isthmus of Panama (Lessios 1979; Vawter et al. 1980).

Values of genetic distance ($D$) have sometimes been calibrated to time ($t$), employing the equation

$$D/2t = cn\lambda,$$

where $c$, $n$, and $\lambda$ represent, respectively, the proportion of all nucleotide substitutions which are electrophoretically detectable, the average number of amino acids per polypeptide, and the average rate of substitutions per codon per year. The value of $t$ can then be determined based upon the value of $D$ and estimates of those of $c$, $n$, and $\lambda$. Nei (1975), for example, has argued that $t = 5 \times 10^4D$ for many vertebrate taxa. Even if the values of $cn\lambda$ averaged over many loci remain relatively constant over moderate time intervals and over many taxa, however, the products of reasonable estimates of $c$, $n$, and $\lambda$, whose true values are currently unknown, generate a broad range of possible parameters for estimating divergence times (Nevo et al. 1974). Thus relatively small differences in, for example, the value of $c$ can yield divergence times which differ many-fold.

Previous estimates of divergence time among taxa have been subject to one or more of the sources of error discussed above. Despite these potential errors, rates of gene substitutions averaged over many loci remain relatively constant over moderate time intervals (Fitch and Langley 1976). Thus, if at least one well-established date is known for the sudden emergence of an absolute barrier to gene flow between any two closely related taxa, genetic distances (generated by the same laboratory procedures) between these and other closely related taxa can be calibrated to give approximate estimates of divergence time that are independent of values of $c$, $n$, and $\lambda$. This set of circumstances has not yet been documented for the comparison of any two taxa.

In this paper, the electrophoresis of a large number of proteins is employed to estimate the time of subgeneric divergence of ground squirrels within the genus *Spermophilus*. The genetic distance estimated by Nei's model (1972), that most frequently used for this purpose, is calibrated to time using a well-established estimate for a geological date associated with the sudden formation of an absolute barrier to gene flow between two ground squirrel subspecies. To test the utility of Nei's model, we include a minimum time of divergence of subgenera *Otospermophilus* and *Spermophilus* based on the fossil record.
Material and Methods

Population Sampling

For subspecies comparisons within the subgenus *Otospermophilus*, blood samples were obtained from 69 Beechey ground squirrels (*Spermophilus beecheyi beecheyi*) and 65 Douglas ground squirrels (*S.b. douglasii*) from California between the spring of 1981 and the fall of 1982. Fifty-three of the Beechey ground squirrels were trapped at Stanford and 16 were trapped at Sunol, about 18 km east of Stanford. Of the 65 Douglas ground squirrels, 36 were trapped near Winters, while the remaining 29 were trapped in Davis, about 18 km east of Winters. After capture, each animal was bled (approximately 3 ml) by cardiac puncture, then released as closely as possible to the location where it was trapped. These samples were compared with those from eight Arctic ground squirrels (*S. paryii ablusus*), which are members of the subgenus *Spermophilus*. These animals were trapped during the summers of 1980 and 1981 along the Denali highway about 100 miles east of Mt. McKinley in central Alaska.

The California field sites were selected because they are all approximately equidistant from the Sacramento River and Delta drainage system, which provides the absolute boundary for the ranges of Beechey and Douglas ground squirrels. Since its origin, the extent of this barrier has fluctuated considerably, with rising sea levels during interglacial periods and with increased winter precipitation during glacial periods (Johnson 1977). Prior to the Delta levee construction, the high tide shoreline of the Delta tidal marsh extended as far east as the city of Sacramento (Atwater 1979). Complementing the tidal marsh as an effective barrier to gene flow during the Holocene, periodic flooding of the Sacramento/San Joaquin River basins produced a temporary lake filling most of the Great Valley (see Brewer 1949). Such periodic flooding would have decimated Beechey and Douglas populations expanding eastward and northward into those river basins and prevented the formation of an intermediate zone of hybridization through which gene flow could occur. Even today, ground squirrels living near the Sacramento River floodplain are periodically devastated by high waters.

Beechey and Douglas ground squirrel subspeciation north and south of the contemporary Sacramento/San Joaquin Delta and San Francisco Bay was initiated by the sudden onset of Great Valley drainage through this riverine system. Radiometric analyses of Rockland ash in the Merced Formation indicate Great Valley drainage into the San Francisco Bay prior to a zircon fission-track date of 0.45 Myr ago (Meyer et al. 1980). Deposits of river sediment underlying the Rockland ash indicate even earlier drainage which, according to Andrei Sarna-Wojcicki (personal communication, U.S. Geological Survey), occurred when the expanding Corcoran Lake suddenly overflowed into the San Francisco Bay during an episode of Nebraskan deglaciation.

Deposits of Bishop Tuff from the massive eruption of the Long Valley caldera (Bailey et al. 1976) covered the top of the Corcoran Clay near Bakersfield shortly after the Corcoran Lake had completely drained (Davis et al. 1977), rerouting the Sacramento River drainage into San Francisco Bay. The accepted K-Ar age of the Bishop Tuff is 0.725 ± 0.015 Myr ago (Dalrymple 1980), thus providing the calibration for the sudden onset of complete and permanent reproductive isolation of Beechey and Douglas ground squirrels. While allopatric speciation could theoretically predate 0.725 Myr ago, hybridization of these two subspecies in captivity...
and among other subspecies of *S. b.* in other regions of California, where their ranges converge at less effective geographic barriers, suggests that allopatric speciation is very unlikely to have occurred.

Ground squirrel fossils from the McKay Reservoir near Pendleton, Oregon (Shotwell 1956; Black 1963, 1972), have been classified as members of subgenera *Spermophilus* (Arctic ground squirrel subgenus) and *Otospermophilus* (California ground squirrel subgenus). The fossils of subgenus *Spermophilus* are the earliest fossils which have been reliably assigned to that subgenus (Black 1963). The Hemphillian deposits within which these fossils were found characterize those of Asia and Europe, indicating continental migrations across the Bering land bridge (see Repenning 1980) constrained by sea level changes between 5.3 and 6.6 Myr ago (Vail and Hardenbol 1979). Thus, the ground squirrel fossils are no younger than 5 Myr (Repenning, personal communication, U.S. Geological Survey), the minimum possible date for subgeneric divergence of Arctic and California ground squirrels.

**Electrophoresis**

After being collected into tubes containing anticoagulant (acid-citrate-dextrose), all blood samples were immediately refrigerated in wet ice. Upon return from the field, samples were centrifuged and the plasma removed, aliquoted into small samples, and frozen at −70°C until used. Red cells were washed three times with 0.85% saline, lysed with an equal volume of distilled water, extracted with toluene, then similarly aliquoted into small samples and frozen until used for analyses described below.

Samples were screened for electrophoretically detectable variation in proteins representing 37 genetic loci. Transferrin (Tf), albumin (Al), alpha-2-macroglobulin (Mac), and NADH (Dia 1) and NADPH (Dia II) diaphorases (EC 1.6.2.2) were screened on 7.5% polyacrylamide gels and glucose phosphate (phosphohexose) isomerase (GPI; EC 5.3.1.9) was screened on starch using methods described by Smith (1980). Carbonic anhydrase (CA I and CA II; EC 4.2.1.1) and esterase D (Est D; EC 3.1.1.1) were screened on starch using the method of Hopkinson et al. (1974), which employs the fluorogenic substrates 4-methylumbelliferyl acetate and fluorescein diacetate for staining CA I (and Est D) and CA II, respectively. Both haptoglobin (Hp) and ceruloplasmin (Cp; EC 1.16.3.1) were run and stained on the same acrylamide gel as described by McCombs and Bowman (1969). Adenosine deaminase (ADA; EC 3.5.4.4; Spencer et al. [1968]), isocitrate dehydrogenase (IDH; EC 1.1.1.42; Chen et al. [1972]) and alpha,-antitrypsin (AAT; Fagerhol [1972]) were all screened on starch gels using methods previously described. The phosphoglucomutase (PGM I and II; EC 2.7.5.1), peptidase (Pep A, B, C and D; EC 3.4.11 and EC 3.4.13.9), lactate dehydrogenase (LDH A and B; EC 1.1.1.27), 6-phosphogluconate dehydrogenase (6-PGD; EC 1.1.1.44), fumarase (Fum; EC 4.2.1.2), glucose-6-phosphate dehydrogenase (G-6-PD; EC 1.1.1.49), malate dehydrogenase (MDH; EC 1.1.1.37), leucine amino peptidase (LAP; EC 3.4.11.1), adenylate kinase (AK; EC 2.7.4.3), hexokinase (Hk; EC 2.7.1.1), the serum and red cell esterases (Est, and Est,; EC 3.1.1.1) and superoxide dismutase (SOD A and B; EC 1.15.1.1) proteins were screened using standard starch gel electrophoresis with histochemical staining methods (Harris and Hopkinson 1976). Hemoglobin (Hb) was screened by cellulose acetate electrophoresis using a Tris-Glycine-EDTA buffer, pH 9.2 (Smith and Ferrell 1980), and, subsequently, using
the Tris-Borate-EDTA buffer containing urea and 2-mercaptoethanol described by Ueda and Schneider (1969). Third complement component (C3) and properdin factor B(Bf) were screened by prolonged agarose gel electrophoresis followed by immunofixation staining (Alper et al. 1971; Ziegler et al. 1975), and catalase (Cat; EC 1.11.1.6) was screened by the method described by Shaw and Prasad (1970).

Estimating Divergence Time from Genetic Distances

Nei's (1972) model was employed for estimating the average genetic distance, in units of gene substitution (D), separating any two of the five population samples of ground squirrels. By this model

\[ D = -\ln I, \]

where

\[ I = \frac{J_{xy}}{(J_x J_y)^{1/2}}. \]

\( J_x, J_y, \) and \( J_{xy} \) are the values of \( \Sigma_i x_i^2, \Sigma_i y_i^2, \) and \( \Sigma_i x_i y_i, \) respectively, which are the sums of the squared frequencies (and cross products of frequencies) of the \( i \)th allele averaged over all loci for two populations, \( X \) and \( Y. \) Nei (1972) has shown that such estimates of \( D \) are linear in \( t, \) as illustrated in equation (1). The method of Nei (1978a) which adjusts estimates of \( D \) to account for small sample sizes was not used. With an average heterozygosity below 0.03 and a high \( D \) value, even the sample of only eight Arctic ground squirrels is more than adequate for minimizing genetic variance caused by polymorphic loci shared with samples of California ground squirrels (Gorman and Renzi 1979).

Nei et al. (1976) have shown that variation in the rate of amino acid substitution among loci follows a gamma distribution, and Nei (1978b) has recommended, on this basis, a modification of the estimate of \( D. \) Thus, we have also estimated values of

\[ D_v = \frac{1 - I}{I}, \]

which corrects for nonconstancy in the rates of gene substitution among loci.

Results

Our examination of genetic variation at 37 loci between the two subspecies of California ground squirrels revealed substantial within-species variation. Nearly one-third of these loci revealed polymorphic variation. When the Arctic ground squirrels are also considered (see table I), 22 loci were polymorphic within our entire sample of the genus *Spermophilus*.

Table 2 gives the values of \( D \) and \( D_v \) for all comparisons using Nei's models. The date of 0.725 Myr ago for subspecific divergence was used to calibrate the subgeneric divergence time (\( t^{\text{subgen}} \)) based on the average proportionality generated by paired comparisons of all relevant within- and between-species \( D \) (and \( D_v \)) values. Nei's model for calculating \( D \) estimated the date for subgeneric divergence at 4.0 Myr ago. When Nei's \( D_v \) values were employed that correct for nonconstancy of rates of amino acid substitutions among proteins, a subgeneric divergence time of about 5.2 Myr ago was obtained. Considering both \( D \) and \( D_v \), the average values for subspecific divergence were 0.103 (± 0.017) and 0.109 (± 0.019), respectively. Similarly, the average \( D \) and \( D_v \) values for subgeneric divergence were 0.563 (± 0.052) and 0.759 (± 0.091), respectively (see table 2).
Table 1
Gene Frequencies for 22 Proteins in the Groups of Genus Spermophilus Which Exhibit within-Genus Genetic Variation*

<table>
<thead>
<tr>
<th>Locus</th>
<th>Alleles</th>
<th>S. b. douglasii</th>
<th>S. b. beecheyi</th>
<th>S. p. abiusus (Arctic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF A</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>1.000</td>
</tr>
<tr>
<td>TF B</td>
<td>.000</td>
<td>.017</td>
<td>.896</td>
<td>.563</td>
</tr>
<tr>
<td>TF C</td>
<td>1.000</td>
<td>.983</td>
<td>.104</td>
<td>.437</td>
</tr>
<tr>
<td>GPI 1</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>1.000</td>
</tr>
<tr>
<td>GPI 2</td>
<td>.972</td>
<td>.966</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>GPI 3</td>
<td>.028</td>
<td>.034</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>PGI 1</td>
<td>.889</td>
<td>.966</td>
<td>1.000</td>
<td>.938</td>
</tr>
<tr>
<td>PGI 2</td>
<td>.111</td>
<td>.034</td>
<td>.000</td>
<td>.062</td>
</tr>
<tr>
<td>PGI 1</td>
<td>.097</td>
<td>.121</td>
<td>.453</td>
<td>.406</td>
</tr>
<tr>
<td>PGI 2</td>
<td>.903</td>
<td>.879</td>
<td>.547</td>
<td>.594</td>
</tr>
<tr>
<td>Hb S'</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Hb S</td>
<td>.264</td>
<td>.242</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Hb F</td>
<td>.319</td>
<td>.293</td>
<td>.849</td>
<td>.750</td>
</tr>
<tr>
<td>Hb SF</td>
<td>.403</td>
<td>.465</td>
<td>.151</td>
<td>.250</td>
</tr>
<tr>
<td>Al 1</td>
<td>.014</td>
<td>.000</td>
<td>.000</td>
<td>.031</td>
</tr>
<tr>
<td>Al 2</td>
<td>.319</td>
<td>.357</td>
<td>1.000</td>
<td>.969</td>
</tr>
<tr>
<td>Al 3</td>
<td>.681</td>
<td>.643</td>
<td>.000</td>
<td>1.000</td>
</tr>
<tr>
<td>AAT 1</td>
<td>.014</td>
<td>.000</td>
<td>.991</td>
<td>1.000</td>
</tr>
<tr>
<td>AAT 2</td>
<td>.972</td>
<td>1.000</td>
<td>.009</td>
<td>.938</td>
</tr>
<tr>
<td>IDH 1</td>
<td>.000</td>
<td>.000</td>
<td>.642</td>
<td>.156</td>
</tr>
<tr>
<td>IDH 2</td>
<td>1.000</td>
<td>.983</td>
<td>.158</td>
<td>.844</td>
</tr>
<tr>
<td>IDH 3</td>
<td>.000</td>
<td>.017</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Pep C 1</td>
<td>.000</td>
<td>.000</td>
<td>.019</td>
<td>.000</td>
</tr>
<tr>
<td>Pep C 2</td>
<td>.972</td>
<td>1.000</td>
<td>.887</td>
<td>.012</td>
</tr>
<tr>
<td>Pep C 3</td>
<td>.028</td>
<td>.094</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Al 1</td>
<td>.000</td>
<td>.000</td>
<td>1.000</td>
<td>.000</td>
</tr>
<tr>
<td>Al 2</td>
<td>1.000</td>
<td>1.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Cat 1</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.000</td>
</tr>
<tr>
<td>Cat 2</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Dia 1</td>
<td>.000</td>
<td>.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Dia 2</td>
<td>1.000</td>
<td>1.000</td>
<td>.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Est 1</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.000</td>
</tr>
<tr>
<td>Est 2</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>1.000</td>
</tr>
<tr>
<td>AK 1</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.000</td>
</tr>
<tr>
<td>AK 2</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Est 1</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.000</td>
</tr>
<tr>
<td>Est 2</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Hb 1</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.000</td>
</tr>
<tr>
<td>Hb 2</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Mac 1</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.000</td>
</tr>
<tr>
<td>Mac 2</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>1.000</td>
</tr>
<tr>
<td>CA 1</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.000</td>
</tr>
<tr>
<td>CA 2</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>1.000</td>
</tr>
<tr>
<td>CA 1</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.000</td>
</tr>
<tr>
<td>CA 2</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Dia 1</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.000</td>
</tr>
<tr>
<td>Dia 2</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Pep B 1</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.000</td>
</tr>
<tr>
<td>Pep B 2</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Pep D 1</td>
<td>.000</td>
<td>.000</td>
<td>.125</td>
<td>.000</td>
</tr>
<tr>
<td>Pep D 2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.875</td>
</tr>
<tr>
<td>Pep D 3</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

* Proteins monomorphic for the same gene in all groups of genus Spermophilus studied were C3, Bf, Est D, 6-PGD, Fum, MDH, LDH A, LDH B, Pep A, ADA, Hk, LAP, G-6-PD, SOD A, SOD B, and Cp. See Electrophoresis section of Material and Methods for meaning of abbreviations.

b This duplicate albumin locus was absent in all Arctic ground squirrels studied and was excluded from the genetic distance analysis.

### Table 2
Values of $D$ and $D_{r}$ for all Possible Comparisons between Subspecies and Subgenera using the Method of Nei

<table>
<thead>
<tr>
<th></th>
<th>$D^{bi}$</th>
<th>$D^{bi}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between subspecies:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Davis-Stanford</td>
<td>.117</td>
<td>.125</td>
</tr>
<tr>
<td>Davis-Sunol</td>
<td>.088</td>
<td>.092</td>
</tr>
<tr>
<td>Winters-Stanford</td>
<td>.119</td>
<td>.126</td>
</tr>
<tr>
<td>Winters-Sunol</td>
<td>.089</td>
<td>.093</td>
</tr>
<tr>
<td>Mean</td>
<td>.103</td>
<td>.109</td>
</tr>
<tr>
<td>SD</td>
<td>.017</td>
<td>.019</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$D^{a}$</th>
<th>$D^{a}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between subgenera:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Davis-Arctic</td>
<td>.518</td>
<td>.679</td>
</tr>
<tr>
<td>Winters-Arctic</td>
<td>.521</td>
<td>.684</td>
</tr>
<tr>
<td>Stanford-Arctic</td>
<td>.594</td>
<td>.812</td>
</tr>
<tr>
<td>Sunol-Arctic</td>
<td>.620</td>
<td>.859</td>
</tr>
<tr>
<td>Mean</td>
<td>.563</td>
<td>.759</td>
</tr>
<tr>
<td>SD</td>
<td>.052</td>
<td>.091</td>
</tr>
</tbody>
</table>

$t^{a}$ (Myr ago) $= 4.04 + .68 \times 5.16 + .98$

where $n$ and $m$, respectively, represent the sum of the number of $D^{bi}$ and $D^{a}$ (or $D^{bi}$ and $D^{a}$) values ($n = 4 = m$) shown above.

### Discussion

Nei's models (1972, 1978b) for estimating genetic distance were tested to determine their ability to predict ground squirrel subgeneric divergence within the genus *Spermophilus*. Using these models, we find the average proportionality between all relevant paired comparisons of $D$ generated subgeneric divergence estimates which underestimated by 19% the 5.0-Myr minimum date of divergence set by the fossil record. The corresponding values of $D_{r}$, however, did approximate the 5-Myr minimum date for subgeneric divergence.

Black (1963) postulates that subgeneric divergence occurred in the Hemphillian between 5 and 8 Myr ago. In agreement, Charles Repenning of the USGS recommends 6.5 Myr ago as a reasonable estimate for the onset of subgeneric divergence. His estimate is coincident with the onset of late Miocene cooling conditions (see Bandy 1968; Vail and Hardenbol 1979) that could have acted as major sources of selection, shifting the subgenus *Spermophilus* into its present alpine and cold-adapted niche.

Thus it is possible that the time of subgeneric divergence based upon Nei's $D_{r}$ values could be underestimated. Nozawa et al. (1982) have reported a more serious underestimate when $D_{r}$ is employed for predicting well-established fossil dates for divergence between primate taxa. The underestimate discussed by Nozawa et al. (1982), and the possible underestimate considered here, could result
from secondary gene substitutions leading to the cancellation of prior charge changes. Nei and Chakraborty (1973) have shown that error resulting from such charge reversals becomes increasingly significant as $D_v$ exceeds 0.5.

Alternatively, the proliferation of fossils of advanced members of subgenus *Spermophilus* after, but not before, the late Hemphillian (Black 1963) could indicate that subgeneric divergence was a punctuated evolutionary event. If the latter view were true, calibration of Nei's $D_v$ values for subspecific divergence at 0.725 Myr ago provides an excellent estimate of the time of subgeneric divergence.

Using this date for subspecific ground squirrel divergence, we find $t = 6.8 \times 10^6 D_v$. This calibration factor is 36% higher than that estimated by Nei (1975) for other taxa and suggests an electrophoretically detectable substitution rate ($cn\lambda = D/2t$) of about $7.4 \times 10^{-8}$ per codon per year. While this difference suggests that evolutionary rates in the genus *Spermophilus* are slower than those of other taxa within which $5 \times 10^6 D$ is considered an appropriate calibrator, extreme caution should be practiced in the use of such indirect calibrating factors. Nozawa et al. (1982) have pointed to serious discrepancies resulting from use of Nei's indirect calibrating factor ($t = 5 \times 10^6 D_v$) for estimating time of divergence among primates, whose evolutionary rates are generally regarded to be slower, not faster, than those of other vertebrate taxa. These discrepancies could result from the assumption that the value of $c$ is much closer to its theoretically maximum value than it really is. It is now known that the value of $c$ varies substantially among laboratories because of differences in laboratory techniques (Bruce and Ayala 1979; Nozawa et al. 1982), and, therefore, comparisons of evolutionary rates generated for even the same species or subspecies among different laboratories might be invalid. Moreover, the substantial increases in genetic variation revealed by sequential gel electrophoresis (Singh et al. 1976; Ramshaw et al. 1979), compared with that resolved by conventional starch gel methods, suggest that the value of $c$ associated with most studies of molecular evolution might be far below its theoretical maximum. If, for example, $n = 400$ and $\lambda = 10^{-9}$ holds true for a broad range of vertebrate taxa (e.g., see Fitch and Langley 1976; Coates and Stone 1981), our estimate of $D_v/2t = cn\lambda = 7.4 \times 10^{-8}$ for the genus *Spermophilus* is consistent with a $c$ value of 0.185, a value about 40% lower than that generally presumed (e.g., King and Wilson 1975; Nei 1975). If the value of $c$ is this low for other electrophoretic comparisons of primate species, then estimates of divergence times based upon Nei's calibrator (e.g., Nozawa et al. 1982) might have been seriously underestimated. If the evolutionary rates of electrophoretically detectable primate polymorphisms are truly slower than those of other taxa, as is widely believed, then much more serious underestimates of $t$ will have been made.

It has been hypothesized that the value of $cn\lambda$ for extracellular proteins exceeds by tenfold that for intracellular proteins (King and Wilson 1975; Sarich 1977). In contrast, the values of $cn\lambda$ for divergence among Douglas, Beechey, and Arctic ground squirrels were between $6.0 \times 10^{-8}$ and $9.0 \times 10^{-8}$ for both extracellular and intracellular proteins. Our results, therefore, are consistent with those of others who have found no evidence for a difference in the evolutionary rates of extracellular and intracellular proteins (Bruce and Ayala 1979; Nozawa et al. 1982). It is more likely that presumed differences in evolutionary rates of intracellular and extracellular proteins result from the greater frequency with which the extracellular proteins are separated on polyacrylamide media, which often
provide better discrimination of allelic variants with small charge differences than does starch.

Estimates of genetic distances among subspecies and subgenera of the genus *Spermophilus* calibrated with the well-established geological date of 0.725 Myr ago for the initiation of Great Valley drainage through San Francisco Bay suggest a date of about 5.16 (+0.98) Myr ago for subgeneric divergence. This date is consistent with the minimum time of divergence estimated from both the fossil record and paleoclimatological events. The calibrating date of 0.725 Myr ago might provide a useful tool for examining evolutionary rates among other closely related vertebrate taxa distributed both north and south of San Francisco Bay. Such studies would permit an assessment of the linearity of $D$, among widely divergent taxa, especially those differing in generation length, and could produce estimates of dates of geological events other than those considered here.

**Acknowledgments**

We acknowledge support from National Institutes of Health grants PHS NO1-HD-9-2828 and RR00169 and Faculty Research Grant D-1343 to D. G. Smith and National Science Foundation grant BNS-7906843 and Faculty Research Grant D-922 to R. G. Coss. We thank Les Becker, Naomie Poran, and, in particular, Becky Rolfs for technical assistance. We also thank Drs. Charles A. Repenning and Andrei M. Sarna-Wojcicki from the U.S. Geological Survey for valuable consultation concerning fossil and geological time markers, respectively.

**LITERATURE CITED**


WALTER M. FITCH, reviewing editor

Received July 25, 1983; revision received November 28, 1983.
The Relationship between Codon Boundaries and Multiple Reading-Frame Preferences: Coding Organization of Bacterial Insertion Sequences

David J. Galas and Temple F. Smith
University of Southern California

Theoretical considerations have shown that the five possible overlapping reading-frame configurations differ significantly in their coding flexibility and thus in their information content (Siegel and Fitch 1980; Smith and Waterman 1980). Contrary to expectation, the overlapping frame configuration allowing the greatest coding flexibility is rarely seen, whereas one of the most constraining is common. We point out here that this overlapping reading-frame paradox and an observed but unexplained preference in coding regions for a pyrimidine-purine at codon boundaries (Shepherd 1981; Jones and Kafatos 1982; Smith et al. 1983) are intimately linked. The codon boundary preference, which may be related to translation efficiency or accuracy, places constraints on the evolution of overlapping coding regions. These considerations may help identify actual coding regions in DNA sequences. We have analyzed five sequenced (enteric) bacterial insertion sequences for codon boundary incidences and reading-frame configurations and find that they are consistent with these proposed constraints.

In several bacteriophage and eucaryotic virus genomes and in many bacterial insertion sequences, it has been observed that potential protein encoding regions overlap, sometimes for a significant fraction of their length (Blattner et al. 1974; Sanger et al. 1977; Szybalski 1977; Gingeras et al. 1982). The proteins that are encoded by the insertion sequences are probably involved in the processes by which these elements transpose. Rather simple theoretical considerations allow the identification of the overlapping reading frames with the minimum and maximum constraint on the choice of amino acids in these sequences (Siegel and Fitch 1980; Smith and Waterman 1980). Because the middle base of the codon is the most determinate of the encoded amino acid and its chemical/structural properties, whereas the third base is the least determinate (Dickerson and Geis 1969, p. 23), it follows that the overlapping frame configuration that places the second base opposite the third base is, on the average, less constrained in coding for amino acids than configurations that put the second opposite the second, or the second

1. Key words: codon boundaries, insertion sequences, overlapping coding frames.

Address for correspondence and reprints: David J. Galas, Molecular Biology, University of Southern California, Los Angeles, California 90089.

© 1984 by The University of Chicago. All rights reserved.
Multiple Reading Frames in Insertion Sequences 261

opposite the first. The quantitative measure of this flexibility is the information content (Siegel and Fitch 1980; Smith and Waterman 1980). These frame configurations are depicted in table 1. Detailed analyses show that the expected information contents of the least and most constrained overlapping configurations differ by a factor of more than three (Smith and Waterman 1980). Since the least constrained is rarely found, there must be some additional considerations that override this apparent advantage. For example, in the transposable genetic element called IS5 (insertion sequence), the frame configuration that places the second bases of the codons opposite one another apparently encodes the observed proteins (Rak et al. 1982). In fact, in three other IS elements the only available large open reading frames also have this configuration (Rak et al. 1982; Kröger, personal communication), as shown in figure 1.

An independent observation on codon usage in nonoverlapping, as well as in overlapping, reading frames is the statistically significant preference for codons ending in a pyrimidine nucleotide and/or beginning with a purine (Schoner and Kahn 1981; Shepherd 1981). This codon boundary preference is even stronger in eucaryotic sequences examined than in procaryotic sequences, and the YR/RY ratio is often greater than 2 (Smith et al. 1983). It has been suggested variously, and, in our opinion, unconvincingly, that this preference may be a statistical fossil of the ancient genetic code structure (Shepherd 1981, 1982; Eigen 1978) or a result of protein structure constraints (Grantham et al. 1980). Ikemura (1981) has pointed out that codon usage is also correlated with tRNA abundance. It seems more likely than the statistical fossil argument, as argued below, that the boundary preference is related to the efficiency or accuracy of the modern translation process itself (see also Ikemura 1981). Whatever the reason for the observed preference, clearly it can only be absolutely maintained in the overlapping reading-frame configuration in which the first bases of the codons on one DNA strand are always opposite the third bases of codons on the other. This, however, is the second most constrained overlapping frame configuration of the five (Siegel and Fitch 1980; Smith and Waterman 1980), with the middle codon bases always opposite one another (3 in table 1).

However, the codon boundary preference is just that, merely a preference

Table 1

<table>
<thead>
<tr>
<th>Overlapping Frame Configurations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1 (Most flexible)</td>
</tr>
<tr>
<td>2 (Least flexible)</td>
</tr>
<tr>
<td>3 (Next most flexible)</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

Note.—The various possible configurations for overlapping coding regions, with the implied constraints on the nature of the codon boundaries. The numbers indicate the first, second, and third positions within a codon whose translational directions are indicated by the arrows. The purines (R), pyrimidines (Y), and the unconstrained base (N) are shown in the overlapping frames (1–5) as they would be constrained to appear by RNY codons shown in the reference frame (0).
Fig. 1.—Open reading frames in several insertion sequences. The open reading frames of more than 50 codons are shown between AUG or GUG codons and terminator codons. The sequence of IS1 was from Ohtsubo and Ohtsubo (1978). There are minor differences between the sequence of these workers (IS/R) and that reported by Johnsrud (1978), but none that changes the reading frames. The sequences for the other elements were taken from: IS903, Grindley and Joyce (1980); IS5, Schoner and Kahn (1981); Kröger and Hobum (1982); IS4, Klaer et al. (1980); IS2, Ghosal et al. (1979). The numbers at the ends of the arrows indicating the length and direction of the reading frames are the coordinates of the start and stop codons (first and last, respectively) as measured from the end of the element identified as the left end in reported sequences. The numbers in parentheses indicate the phase of the reading frame so that two open frames having the same numbers are in phase, even if they are read in opposite directions. The dashed arrow in IS903 indicates an open reading frame with a high YR:RY that begins with a GTA codon. The dashed arrow in IS2 represents an open frame longer than 50 codons, which does not terminate within the element.

which is a codon usage preference in part 2 and one for which there are a few known exceptions (Smith et al. 1983), notably in the genomes of the mammalian mitochondria. Thus it is of interest to ask, If the first codon position is a purine and the third is a pyrimidine with some given probability, can the order of the different overlapping reading-frame configurations be changed? Calculations were performed by the methods described previously (Smith and Waterman 1980), using conditional probabilities for neighboring codon occurrences but with the added constraint of unequal probabilities for purines and pyrimidines in the first and third positions. The differences in information content (between 1 and 3 in table 1) are significantly smaller than without this constraint, but no inversion in order is possible for any compatible values of the unequal probabilities. Thus, it is not a clear question of information capacity or coding flexibility. However, to maintain a codon boundary preference in both frames in configurations 1, 2, 4, or 5 (table 1) requires the imposition of asymmetric constraints on the middle base position.

2. For the five fourfold and three sixfold degenerate codon sets, simple preferential codon usage could provide pyrimidine preference in third bases, but only in the case of Ser and Arg could synonymous codon usage provide a purine preference in first bases as well.
for the two overlapping coding frames. It is only in configuration 3 that these asymmetric constraints can be avoided. In this case, the boundary preference in one frame is reflected in the other frame with no additional constraints imposed. In this sense it is the least constrained configuration, when the codon boundary preference is imposed. (The full implications of the constraints on position-dependent probabilities are complex but not very enlightening for the present discussion, so they will not be elaborated here [Galas and Smith, unpublished].)

In the bacterial insertion sequences, and in IS5 (Engler and Van Bree 1983) in particular, we have an opportunity to examine several strongly overlapping reading frames (Blattner et al. 1974; Szybalski 1977; Kröger, personal communication). There are two protein-encoding reading frames in IS5 (Rak et al. 1982). The data in table 2 for these two overlapping reading frames strongly support the idea that the "choice" of frame overlap configuration was due to the need to maintain the codon boundary preference. In fact, in these particular sequences the enhancement in the YR over RY codon boundary preference in the overlapping region suggests that the statistical requirement for ending with Y and beginning with R may be independent. In the largest of the IS5 protein coding sequences, the nonoverlapping region has a YR/RY ratio of only 1.3, whereas the overlapping region has a value exceeding 2.0 (see table 2). The apparent requirement cannot be simply to have an average preference for YR, since the presumptive 30% preference in the larger frame automatically exists in the other, given the overlapping configuration between these two coding frames. Thus, a possible implication of this particular frame relationship is that the YR preference at each codon boundary is independently determined and therefore the observed preference in the overlap is multiplicative. Note again that all other frame relationships impose on one or the other coding sequence the opposite preference. We have chosen to use the ratio YR/RY as the indicator for codon boundary preference because this parameter has the important property of being independent of base composition.

As pointed out by Rak et al. (1982) and others (Klaer et al. 1980; Kröger, personal communication), many insertion sequences have coding frames arranged in a characteristic manner—a large, open frame coding in one direction and a significantly shorter frame coding in the opposite direction, with frame configuration 3 in table 1. In figure 1, four such elements with their open reading frames (criteria are in the figure legend) are shown (IS903, IS5, IS2, IS4). In spite of their structural similarity, it should be noted that there is no obvious DNA or amino acid sequence homology among these open frames and their potential proteins (a very weak similarity of doubtful significance exists between IS2 and IS4; data not shown). The first element shown in the figure, IS1, is peculiar in this respect by not having a single, long, open frame. It is both the shortest element of those shown here and the most complex in potential coding frame arrangements. It is like the others, however, in having two opposite-strand, aligned reading frames (3 in table 1) in the left half of the element (frames designated A and D).

When we examine the codon boundary statistics of the open reading frames in all of the elements of figure 1 (table 2), we find that the four elements have preferences for YR at the codon boundaries in the long, open frames and at least one of the opposite-strand, shorter frames. In the region of overlap between opposite-strand, in-frame reading frames, this codon boundary preference is often significantly enhanced, as noted above for the known protein-coding frames in
Table 2
Codon Boundaries in the Known Protein-coding Regions of IS5, IS4, and the Potential Frames in Other Elements

<table>
<thead>
<tr>
<th>IS Element Open Reading Frame</th>
<th>Boundary Ratio</th>
<th>Frame Length in Codons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>YR/RY</td>
<td></td>
</tr>
<tr>
<td>IS1:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>2.8</td>
<td>91</td>
</tr>
<tr>
<td>B</td>
<td>.6 (.7)</td>
<td>125 (167)*</td>
</tr>
<tr>
<td>C</td>
<td>2.1</td>
<td>96</td>
</tr>
<tr>
<td>D</td>
<td>2.6</td>
<td>71</td>
</tr>
<tr>
<td>E</td>
<td>.8</td>
<td>89</td>
</tr>
<tr>
<td>F</td>
<td>1.0</td>
<td>70</td>
</tr>
<tr>
<td>IS2:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>1.7</td>
<td>307</td>
</tr>
<tr>
<td>S</td>
<td>1.9</td>
<td>114</td>
</tr>
<tr>
<td>S</td>
<td>2.0</td>
<td>90</td>
</tr>
<tr>
<td>IS903:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>1.5</td>
<td>326</td>
</tr>
<tr>
<td>S</td>
<td>2.4 (2.5)</td>
<td>116 (108)*</td>
</tr>
<tr>
<td>S</td>
<td>1.8</td>
<td>90</td>
</tr>
<tr>
<td>S</td>
<td>1.0</td>
<td>59</td>
</tr>
<tr>
<td>S</td>
<td>1.0</td>
<td>52</td>
</tr>
<tr>
<td>IS4:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>1.2</td>
<td>315</td>
</tr>
<tr>
<td>S</td>
<td>1.6*</td>
<td>143</td>
</tr>
<tr>
<td>IS5:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>1.3</td>
<td>442</td>
</tr>
<tr>
<td>S</td>
<td>1.5</td>
<td>131</td>
</tr>
<tr>
<td>S</td>
<td>.9</td>
<td>91</td>
</tr>
</tbody>
</table>

NOTE.—The "small" coding frame in IS5 (S,) is completely within the "large" frame (L). The numbers refer to the ratio of the number of boundaries found with the pyrimidine-purine (YR) character to those with the purine-pyrimidine (RY) character. The reading frames indicated in the table are those shown in fig. 1. Ohtsubo et al. 1981 have suggested that A and B in IS1 are real coding regions by comparison of these open frames with a closely related insertion sequence. The numbers in the figure may be compared with the expected value of YR/RY if codon usage were completely random, i.e., all sense codons occur with equal probability (this is, of course, highly unrealistic). In this case the ratio is expected to be 1.2.

* As in fig. 1, only open reading frames beginning with ATG or GTG and containing at least 50 codons were considered.

* These open reading frames have several possible starts. B in IS1 has one further upstream than that shown in fig. 1 and many downstream.

* This frame actually begins with a GTA codon, but it is included here because the YR/RY ratio is quite high.

* This particular reading frame, while having a YR/RY boundary ratio of greater than 1.5, does not show YR as the maximum. YY is the slightly more frequent boundary (by three codons).

IS5. With respect to IS1, we find that A and D have a very strong codon boundary preference for YR (YR/RY ratio greater than 2.5).

A particularly interesting case is encountered for open reading frame B of IS1, which has been implicated as an actual coding frame by Ohtsubo et al. (1981) for other reasons (called InsB in their paper). Here the longest open frame, and thus perhaps the one most strongly expected to encode a real protein, does not display the novel coding region preference (this point has also been recently noted by Iida et al. [1983]). In fact, it displays the opposite preference, whereas the open frame most overlapping B, C, displays a strong YR preference (YR/RY of 2.1). C is not in configuration 3 with respect to B. Note that for two overlapping
reading frames with the codons out of phase, only one can be expected to display a YR boundary preference. In this case it is C rather than B. We have also calculated the third-order Markov correlations\(^3\) for the central base of the codons in B (Smith et al. 1983) and find that it is consistent with a random distribution of the observed base composition in that position. Frame C not only has the expected YR boundary statistics and reasonable third-order Markov correlations but has a presumptive promoter just upstream. By the above statistical criteria, C seems the more likely actual coding frame. However, if B does encode a functional protein, its unusual character might suggest a different origin or that different selection pressures have been exerted on this protein. From the codon boundary preference in frames A–F in IS1 and the histograms in figure 2, it appears that frames A, C, and D are the most likely to encode proteins. Similarly, it would appear that S\(_1\) and S\(_1\) in IS5 and S\(_2\) in IS4 are unlikely to encode proteins. Note that this is true in spite of the fact that, just like S\(_1\), S\(_2\) in IS4 is in frame (3 in table 1) with the opposite-strand encoded frame L, just like S\(_1\).

Under the assumption that, like small viruses or phages, there is a selective advantage to being highly compact and information efficient, one would expect that some, if not all, such systems will use the least informationally constrained (maximum coding flexibility) overlapping frame configurations (1, 2, or 4 in table 1). Since this is apparently not the case, particularly for IS5 where two proteins have been detected (5), we conclude that the generally observed properties of protein coding regions in DNA represent additional constraints, like the codon boundary preference, that supersede the presumed advantage of higher information density. This, in turn, implies that a maximum length, open reading frame as the principal criterion for identifying probable protein coding regions is likely to be misleading, as we suggest it may be in IS1. It should be noted that under the simplest codon usage model (random occurrence of all sense codons with codons with equal probability), the expected average YR/RY codon boundary ratio in long, open reading frames is 1.22. Such a random model is, of course, highly unrealistic for known coding domains (Grantham et al. 1980; Smith and Waterman 1980).

---

\(^3\) There are three diagnostic characteristics of the vast majority of known coding regions: suppression of terminators, strong third-order Markov correlations, and the YR boundary preference (Smith and Waterman 1980).

---

**Fig. 2.** Histograms of the YR/RY ratio for the codon boundaries of the open reading frames shown in fig. 1. The numbers are shown in col. 2 of table 2. The crosshatched bars are for the actual coding regions known in IS5 (Rak et al. 1982). A protein that is apparently coded for by the large open frame in IS4 has been detected (Trinks et al. 1981); neither of the two short frames has been found to produce detectable amounts. The frames that are above the expected value of 1.2 in their YR/RY ratio are IS1 A, C, D; IS903 L, S\(_1\); IS5 L, S\(_1\), S\(_2\); IS2 L, S; IS4 L, S\(_1\).
Waterman 1980; Ikemura 1981). The statistical significance of the observed ratio is hard to evaluate since such things as known codon usage frequencies predict a large fraction of the observed value but may themselves arise in part from a need to maintain this very ratio.

A critical question is, Are there other possible reasons for the observed reading-frame preference? For example, most of the presumptive IS proteins are very arginine-lysine rich, as might be expected of some DNA binding proteins, and one must ask whether this property might influence the reading-frame overlap configuration. Since the three positive amino acids are encoded primarily by RNR and YNY codons, it can be seen (calculations not shown) that the observed configuration (3 in table 1) is as compatible as some others with regions rich in these codons, and therefore the charge of the protein is probably unimportant to the configuration.

The statistical preference discussed here may be considerably more fundamental than has been suspected. Some recent results could be interpreted to imply that this preference is involved in the translation system’s ability to identify the “proper” open reading frames (fig. 1) or to minimize successive frameshift errors in translation (Weiss and Gallant 1983; Kurland and Gallant, accepted). The study of the effect of surrounding sequences on the efficiency of nonsense suppression has recently revealed that a purine on the 3’ side of the codon being suppressed generally enhances the efficiency of translation by suppressor tRNAs through the nonsense codon (Bossi 1983; Miller and Albertini 1983). The strength of the effect appears to depend on the particular suppressor tRNA.

The codon boundary preference has also been noted in a completely different context in the mutational history of a protein gene family by Jones and Kafatos (Jones and Kafatos 1982). A recent study (Milkman and Crawford 1983) of neutral mutations (no amino acid change) of trp operon genes in strains of *Escherichia coli* reveals a very strong bias for purine to pyrimidine changes in the third base position of the codon, which also suggests the importance of maintaining (or enhancing) the codon YR boundary preference. We must conclude that the codon use reflected in the boundary preference is of more than peripheral interest, since it is manifested not only in the overlapping reading-frame configuration “paradox” discussed above but also in a number of other recent investigations cited above (Bossi 1983; Miller and Albertini 1983; Weiss and Gallant 1983; Kurland and Gallant, accepted). In any case, the understanding of this question will be important in investigating the heavily constrained evolution of multiple coding regions. The IS’s may provide an important system in which to investigate this coding property. Note, for example, that elements like IS903, IS2, IS5, and IS4 (fig. 1) that have a long (presumed real) coding frame taking up most of the length of the element are very limited in what other proteins they can code for. Any other open frame of reasonable length must overlap the longest one, and if the codon boundary preference discussed above is to be maintained, there is only one frame on the opposite strand that can be used for encoding a protein. These considerations suggest that the evolution of a second protein coding gene on the opposite strand of an existing, functional gene probably proceeds by the molding of the preadapted, in-frame codons on the opposite strand (configuration 3, table 1). A coding region is much more likely to evolve in this frame because, whatever the source of the advantage of the Y/R preference, it already exists in this form by virtue of its relationship to its counterpart on the other strand.
Acknowledgments

We thank Drs. M. Waterman, J. Petruska, P. Prentki, and M. Chandler for stimulating discussions. This work was supported by National Institutes of Health grant GM 19036 (D. J. G.) and a System Development Foundation grant (T. F. S.). T. F. S. was visiting professor during the course of this work. (Present address: Department of Physics, Northern Michigan University, Marquette, Michigan 49855.)

LITERATURE CITED


WALTER M. FITCH, reviewing editor

Received August 17, 1983; revision received November 1, 1983.
A mathematical formula for estimating the average number of nucleotide substitutions per site (δ) between two homologous DNA sequences is developed by taking into account unequal rates of substitution among different nucleotide pairs. Although this formula is obtained for the equal-input model of nucleotide substitution, computer simulations have shown that it gives a reasonably good estimate for a wide range of nucleotide substitution patterns as long as δ is equal to or smaller than 1. Furthermore, the frequency of cases to which the formula is inapplicable is much lower than that for other similar methods recently proposed. This point is illustrated using insulin genes. A statistical method for estimating the number of nucleotide changes due to deletion and insertion is also developed. Application of this method to globin gene data indicates that the number of nucleotide changes per site increases with evolutionary time but the pattern of the increase is quite irregular.

Introduction

The evolutionary change of DNA sequences occurs by nucleotide substitution, deletion, and insertion. The change due to nucleotide substitution is measured in terms of the number of nucleotide substitutions per site between two homologous DNA sequences. Several statistical methods for estimating this number have been developed. Unfortunately, however, all of them have some deficiencies. Jukes and Cantor's (1969) method is the simplest one but gives underestimates when the rate of nucleotide substitution is not the same for all nucleotide pairs. Recently, Kimura (1980, 1981), Takahata and Kimura (1981), and Gojobori et al. (1982a) developed new methods for estimating the number of nucleotide substitutions, taking into account unequal rates of substitutions among different nucleotide pairs. However, these methods are all dependent on specific schemes of nucleotide substitutions, and if actual nucleotide substitution does not follow these schemes, the methods are expected to give biased estimates. Furthermore, they

1. Key words: nucleotide substitution, evolutionary distance, unequal substitution rates, deletion, insertion, globin genes, insulin genes.

Address for correspondence and reprints: Dr. Masatoshi Nei, Center for Demographic and Population Genetics, University of Texas at Houston, P.O. Box 20334, Houston, Texas 77225.
are often inapplicable to actual data because of a negative argument in the logarithm of the formula used. In this paper we propose a new method that alleviates some of these deficiencies. We shall also consider the evolutionary changes of DNA arising from deletions and insertions and present a method for measuring the amount of these changes.

**Number of Nucleotide Substitutions**

**Theory**

Consider two homologous nucleotide sequences that diverged from a common ancestral sequence \( t \) years ago. We first consider the case where the rate of nucleotide substitution is the same for all pairs of nucleotides and equal to \( \lambda \) per site per year. The expected number of nucleotide substitutions per site between the two sequences for this case is given by

\[
S = 2ht. \quad (1)
\]

If we know the proportion \( n \) of different nucleotides per site, \( \delta \) can be estimated by

\[
\delta = -\frac{3}{4} \log_e (1 - 4\pi/3), \quad (2)
\]

where \( 0 \leq \pi \leq 3/4 \) (Jukes and Cantor 1969; Kimura and Ohta 1972).

At this point, we note that (2) can be written as

\[
\delta = -b_1 \log_e (1 - \pi/b_1), \quad (3)
\]

where \( b_1 = 1 - \Sigma q_i^2 \). Here \( q_i \) is the equilibrium frequency of the \( i \)th nucleotide \( (i = 1, 2, 3, 4 \) corresponding to the nucleotides A, T, G, C). When the rate of nucleotide substitution is the same for all nucleotide pairs, \( q_i = 1/4 \), so that \( b_1 = 3/4 \). We also note that \( b_1 = 3/4 \) is the maximum value of \( \pi \), which is attained at \( t = \infty \).

Kimura (1980, 1981), Takahata and Kimura (1981), and Gojobori et al. (1982a) have shown that when the rate of nucleotide substitution varies with nucleotide pair, (2) gives an underestimate of \( \delta \). Part of the reason is that in this case the equilibrium value of \( \pi \) is generally smaller than \( 3/4 \). Note that in any scheme of nucleotide substitution the value of \( \pi \) at \( t = \infty \) is given by \( h_1 = 1 - \Sigma q_i^2 \). The value of \( q_i \) can be uniquely determined for any substitution scheme (Tajima and Nei 1982). This suggests that (3) may be used as an estimator of \( \delta \) even for the case of unequal substitution rates. The estimate of \( \delta \) obtained by (3) is always equal to or greater than that obtained by (2).

Equation (3) holds exactly for Tajima and Nei’s (1982) equal-input model of nucleotide substitution with unequal rates. Let \( \lambda_{j} \) be the rate of substitution of the \( j \)th nucleotide for the \( i \)th nucleotide per unit evolutionary time. This unit evolutionary time can be, for example, year, generation, or 1,000 years, depending on the purpose. In the equal-input model, \( \lambda_{j} = a_j \) for all \( i \)'s except for \( \lambda_{j} \). In other words, the rate of substitution of the \( j \)th nucleotide for the \( i \)th nucleotide is the same, irrespective of the \( i \)th nucleotide. Therefore, the substitution rate matrix is given by (A1) in the Appendix, where \( \lambda_{j} = 1 - \Sigma a_i \). Using this substitution rate matrix, one can prove (3), as shown in the Appendix.

In practice, of course, the pattern of nucleotide substitution does not necessarily follow this scheme (see Gojobori et al. 1982b). When the substitution
scheme is different from the equal-input model, (3) is no longer valid, as is clear from the works of Kimura (1980, 1981), Takahata and Kimura (1981), and Gojobori et al. (1982a). In this case, however, a slight modification of (3) gives a quite reliable estimate, as will be shown later by computer simulation. This modification is based on the following observations. (i) In the equal-input model, \( c_{ij} = x_{ij}/(2q_i q_j) \) is constant for all \( i \) and \( j \) (\( i < j \)), where \( x_{ij} \) is the proportion of pairs of nucleotide \( i \) and \( j \) between the two homologous DNA sequences (see Appendix). (ii) Our computer simulations discussed in the next section have shown that when \( c_{ij} \) is not constant, (3) tends to give an underestimate. (iii) In the case of the equal-input model, \( \delta \) can also be estimated by using information on the frequencies of nonidentical nucleotide pairs. Namely,

\[
\delta = -2 \sum_{i=1}^{3} \sum_{j=i+1}^{4} q_i q_j \log_e (1 - c_{ij})
\]

\[
= -b_2 \log_e (1 - \pi/b_2),
\]

where \( b_2 = \pi^2/h \) and

\[
h = \sum_{i=1}^{3} \sum_{j=i+1}^{4} x_{ij}^2/(2q_i q_j)
\]

(see Appendix). When \( c_{ij} \) is not constant, however, (4) tends to give an overestimate of \( \delta \) (results from our computer simulations). These observations suggest that an approximate estimate of \( \delta \) is obtained by

\[
\delta = -b \log_e (1 - \pi/b),
\]

where \( b \) is the average of \( b_1 \) and \( b_2 \) and given by

\[
b = \left( 1 - \sum_{i=1}^{4} q_i^2 + \pi^2/h \right) / 2.
\]

It is desirable to know the accuracy of this formula for various patterns of nucleotide substitution. However, analytical evaluation of the accuracy is not easy, because the mathematical property of the most general substitution scheme requiring 12 parameters has not been studied. We have therefore conducted a computer simulation to examine this accuracy. As will be shown in the next section, this simulation indicates that (6) gives a quite reliable estimate as long as \( \delta \) is smaller than 1. Needless to say, equation (6) holds exactly for the case of equal substitution rates or the equal-input model.

So far we have considered the deterministic change of DNA divergence. In practice, the numbers of nucleotide substitutions are studied by examining a finite number of nucleotides, and thus the estimate \( \hat{\delta} \) of \( \delta \) is subject to sampling error. The sampling variance of \( \hat{\delta} \) obtained from (6) is given by

\[
V(\hat{\delta}) = \left( \frac{\partial \delta}{\partial \pi} \right)^2 V(\pi) + \left( \frac{\partial \delta}{\partial b} \right)^2 V(b)
+ 2 \frac{\partial \delta}{\partial \pi} \frac{\partial \delta}{\partial b} \text{cov}(\pi, b).
\]

(8)

It can be shown that the second and third terms of (8) are very small compared
with the first term unless \( n \) (number of nucleotide pairs examined) is unusually small, say, \( n < 40 \). Therefore, we have (approximately)

\[
V(\delta) = b^2 \pi (1 - \pi) / [(b - \pi)^2 n].
\] (9)

**Computer Simulation**

In this section we shall examine two different aspects of the accuracies of the estimates of \( \delta \) obtained by (3) and (6). One is the effect of deviation of nucleotide substitution from the equal-input model, and the other is the effect of sampling error when a relatively small number of nucleotides are examined. In the study of the former effect we assume that the DNA sequence under investigation is infinitely long.

**Effect of Deviation from the Equal-Input Model of Substitution**

Gojobori et al. (1982a) studied the relative rates of nucleotide substitution among the four nucleotides (A, T, G, C) for three functional genes (\( \alpha \) and \( \beta \) globin genes and ACTH gene) and six pseudogenes (four globin pseudogenes, one Ig \( V \), pseudogene, and one U1 snRNA pseudogene). These relative rates were quite different from the rates expected from any of the mathematical models studied so far. Therefore, it is interesting to know which statistical method gives the best estimate of \( \delta \) when nucleotide substitution occurs according to these observed patterns. We therefore used the nine substitution schemes observed to simulate the evolutionary change of nucleotide sequences. In this simulation we followed Gojobori et al.'s (1982a) method and computed the \( \delta \) values for the nine substitution schemes. That is, the matrix of relative substitution rates \( (p_{ij}; i \neq j) \) was first converted into the matrix of substitution rates \( (\lambda_{ij}) \) corresponding to \( k = \Sigma_i q_i \Sigma_j \lambda_{ij} = 0.0078125 \), where \( k \) is the average number of nucleotide substitutions per unit evolutionary time. The values of \( x_0 \)'s for \( \delta = 0.25, 0.5, 1.0, \) and 2.0 were then obtained by squaring the matrix of substitution rates repeatedly (see Gojobori et al. [1982a] for details). Note that \( \delta = 0.25, 0.5, \) etc. are obtained by squaring the matrix five times, six times, etc. From the values of \( x_0 \)'s, \( \delta \) was estimated by using seven different estimation methods, that is, (a) the Jukes-Cantor (JC) method, (b) Kimura's (1980) two-parameter (2P) method, (c) Kimura's (1981) three-substitution-type (3ST) method, (d) Takahata and Kimura's (1981) (TK) method, (e) Gojobori et al.'s (1982a) (GIN) method, (f) equation (3), and (g) equation (6). The deviation of the estimate from the true value of \( \delta \) was measured by the following bias index:

\[
B = \left[ \sum_{i=1}^{r} (\hat{\delta}_i - \delta)^2 / r \right]^{1/2},
\] (10)

where \( \hat{\delta}_i \) is the estimate of \( \delta \) for the \( i \)th substitution scheme and \( r \) is the number of substitution schemes used. In the present case \( r = 9 \).

The \( B \) values obtained are presented in table 1. It is clear that when \( \delta \) is small, that is, \( \delta \leq 0.5 \), equation (6) gives an estimate of \( \delta \) with the smallest amount of bias, whereas when \( \delta \geq 1.0 \), the TK and GIN methods tend to give a better estimate than equation (6). Equation (3) gives a smaller value of \( B \) than the JC, 2P, and 3ST methods for all values of \( \delta \), but the bias of the estimate obtained by
Table 1

Bias Indices \((B)\) of the Estimates of \(\delta\) Obtained by Seven Different Methods for Various Schemes of Nucleotide Substitution

<table>
<thead>
<tr>
<th>True (\delta)</th>
<th>JC</th>
<th>2P</th>
<th>3ST</th>
<th>TK</th>
<th>GIN</th>
<th>(3)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nine substitution schemes observed for actual genes:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.25</td>
<td>.020</td>
<td>.019</td>
<td>.018</td>
<td>.016</td>
<td>.017</td>
<td>.014</td>
<td>.003</td>
</tr>
<tr>
<td>.50</td>
<td>.072</td>
<td>.069</td>
<td>.068</td>
<td>.022</td>
<td>.045</td>
<td>.053</td>
<td>.018</td>
</tr>
<tr>
<td>1.00</td>
<td>.240</td>
<td>.230</td>
<td>.223</td>
<td>.054</td>
<td>.135</td>
<td>.177</td>
<td>.108</td>
</tr>
<tr>
<td>2.00</td>
<td>.728</td>
<td>.698</td>
<td>.675</td>
<td>.381</td>
<td>.361</td>
<td>.514</td>
<td>.449</td>
</tr>
<tr>
<td>Thirty-one substitution schemes artificially generated:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.25</td>
<td>.012</td>
<td>.012</td>
<td>.011</td>
<td>.009</td>
<td>.014</td>
<td>.007</td>
<td>.003</td>
</tr>
<tr>
<td>.50</td>
<td>.047</td>
<td>.046</td>
<td>.045</td>
<td>.027</td>
<td>.029</td>
<td>.026</td>
<td>.011</td>
</tr>
<tr>
<td>1.00</td>
<td>.178</td>
<td>.168</td>
<td>.164</td>
<td>.093</td>
<td>.060</td>
<td>.096</td>
<td>.061</td>
</tr>
<tr>
<td>2.00</td>
<td>.635</td>
<td>.582</td>
<td>.566</td>
<td>.322</td>
<td>.136</td>
<td>.316</td>
<td>.273</td>
</tr>
</tbody>
</table>


Bias indices were computed by (10).

(3) is larger than that of (6). As expected, the JC method gives an estimate of \(\delta\) with the largest bias for all values of \(\delta\).

Since nine substitution schemes would not be sufficient for drawing a general conclusion, we used 31 more different schemes which were generated artificially by using random numbers. Seven substitution schemes were obtained by assuming that each element of the matrix of relative substitution rates \((P_{ij})\) takes one value of 0.001, 0.002, \ldots, 0.009, and 0.01 with equal probability (1/10). The remaining 24 substitution schemes were obtained by assuming that \(P_{ij}\) takes one value of 0.001, 0.002, \ldots, 0.009, and 0.01 with probabilities 0.19, 0.17, 0.15, \ldots, 0.03, and 0.01, respectively. The \(P_{ij}\) matrices thus obtained covered a wide range of substitution patterns. The \(P_{ij}\) matrices were then converted into the substitution rate matrix corresponding to \(k = 0.0078125\). Using these matrices, we again estimated \(\delta\)'s by using the seven statistical methods.

The \(B\) values for these new simulations are given in the lower half of table 1. When \(\delta \leq 0.5\), equation (6) again gives the best result, the \(B\) value being considerably smaller than that for the other methods. When \(\delta \geq 1\), however, the GIN method is superior to (6), though the latter is better than the TK method. Considering this case together with the case of empirical substitution schemes mentioned above, we can conclude that (6) is better than the other methods in estimating \(\delta\) when \(\delta\) is small, whereas the GIN method gives the best result when \(\delta\) is large.

Although our bias index gives the average bias of the estimates of \(\delta\), it does not give information about the direction of the bias. This information is provided in figure 1, where the distribution of \(\hat{\delta}\) is given in relation to \(\delta\) for the four levels of \(\delta\). The distributions of \(\hat{\delta}\) for the 2P and 3ST methods and equation (3) are not given here, because these are apparently inferior to equation (6). It is seen that when \(\delta\) is 0.25 or 0.5, equation (6) gives a very narrow distribution around the true value of \(\delta\). The GIN method gives a mean value of \(\hat{\delta}\) close to the true value,
but the deviation from the true value is often large. However, the JC method almost always gives an underestimate of $\delta$. The TK method also tends to give an underestimate, but the extent of underestimation is not as bad as that of the JC method. When $\delta \geq 1$, however, all methods except the GIN method give underestimates, but the extent of underestimation for equation (6) is small when $\delta = 1$. The GIN method generally gives an average estimate close to the true value of $\delta$ and a small value of $B$, though the $B$ value for the case of $\delta = 1$ is slightly larger than that for equation (6). From figure 1, therefore, we may conclude that equation (6) is superior to the other methods when $\delta \leq 1$, but when $\delta > 1$ the GIN method is probably the best one.

**Sampling Error**

When the number of nucleotides compared is small, the estimates of $q_i$ and $x_{ij}$ may deviate from the expected values by chance, and this deviation is expected to affect the estimate of $\delta$ or produce cases to which equation (6) or other methods are inapplicable because of a negative argument in the logarithm involved. To examine the magnitude of this error, we conducted another computer simulation. In this simulation we considered three different numbers of nucleotides, that is, $n = 50, 144,$ and 500. The latter two numbers were chosen to compare our results with those of Gojobori et al. (1982a). In Gojobori et al.’s computer simulation many inapplicable cases were produced when their six-parameter model of nucleotide substitution was used. Since we were primarily interested in the frequency

---

**FIG. 1.—Distribution of the estimates ($\hat{\delta}$) of $\delta$ obtained by four different methods. I, JC method. II, eq. (6). III, TK method. IV, GIN method. A, B, C, and D represent the cases of $\delta = 0.25, 0.5, 1,$ and 2, respectively. Arrows indicate the locations of the true values of $\delta$. The scale of $\delta$ varies with $\delta$. The total number of observations is 40 in each case.**
of inapplicable cases, we used the same substitution model. The substitution rates used were $\alpha = 0.00125$, $\alpha_i = 0.008$, $\alpha_j = 0.118$, $\beta = 0.005$, $\beta_i = 0.004$, and $\beta_j = 0.0059$ with $k = 0.01$, where the parameters $\alpha$, $\alpha_i$, etc. are identical to those given in Gojobori et al.'s (1982a) table 2. Ancestral sequences of 50, 144, and 500 nucleotides were generated by using pseudorandom numbers. From each of these ancestral sequences, 50 pairs of descendant nucleotide sequences were randomly produced for each of $\delta = 1.0$ and 2.0 by using the method described by Gojobori et al. (1982a). For each pair of descendant sequences, $x_{ij}'s$ were computed, and $q_i = x_{ii} + \Sigma x_{ij}/2$ was obtained. Using these $q_i's$ and $x_{ij}'s$, we estimated $\delta$ by the JC method and equations (3) and (6). In the case of $\delta = 2.0$, the $\delta$ values for $n = 50$ were not computed, since in this case an estimate of $\delta$ is obviously unreliable because of a large sampling error.

The mean ($\bar{\delta}$) and standard deviation ($\sigma_\delta$) of $\delta$ obtained and the frequency of inapplicable cases ($f$) are given in table 2. In this case the values for $n = \infty$, which can be obtained theoretically, are also presented. The JC method again gives underestimates of $\delta$ for both $\delta = 1$ and 2, but there are no inapplicable cases. Equation (3) gives a much better estimate of $\delta$; however, there are a few inapplicable cases. Equation (6) gives an even better estimate of $\delta$ than equation (3), but the number of inapplicable cases is slightly larger than that for (3). Table 3 gives the results obtained by Gojobori et al. (1982a) for the TK and GIN methods. In both methods the frequency of inapplicable cases is very high compared with that of (3) and (6). If we remove inapplicable cases, however, the GIN method gives a relatively good estimate, though the variance is quite large. The TK method also gives a good estimate of $\delta$ when $\delta = 1$ but a serious underestimate when $\delta = 2$. From these results we can conclude that our equations (3) and (6) are less sensitive to sampling error than the TK and GIN methods.

Table 2 includes the observed and expected standard deviations of $\delta$. The observed values were computed from replicate estimates of $\delta$ with the inapplicable cases excluded, whereas the expected values were obtained from (9). If we con-

<table>
<thead>
<tr>
<th>TRUE $\delta$ AND $n$</th>
<th>JC METHOD</th>
<th>EQUATION (3)</th>
<th>EQUATION (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\bar{\delta}$</td>
<td>$\sigma_\delta$</td>
<td>$f$</td>
</tr>
<tr>
<td>1.0:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50......................</td>
<td>.81</td>
<td>.21</td>
<td>.20</td>
</tr>
<tr>
<td>144.....................</td>
<td>.82</td>
<td>.16</td>
<td>.12</td>
</tr>
<tr>
<td>500.....................</td>
<td>.78</td>
<td>.05</td>
<td>.06</td>
</tr>
<tr>
<td>$\infty^b$.............</td>
<td>.79</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>2.0:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>144.....................</td>
<td>1.22</td>
<td>.20</td>
<td>.20</td>
</tr>
<tr>
<td>500.....................</td>
<td>1.22</td>
<td>.12</td>
<td>.11</td>
</tr>
<tr>
<td>$\infty^b$.............</td>
<td>1.20</td>
<td>.00</td>
<td>.00</td>
</tr>
</tbody>
</table>

Note.—$\bar{\delta}$ = average of the estimate ($\bar{\delta}$) of $\delta$, $\sigma_\delta$ = standard deviation of the estimate, $\sigma_\delta$ = expected standard deviation obtained from formula (9), $f$ = proportion of inapplicable cases, and $n$ = number of nucleotide pairs. The number of replications used is 50.

$^a$ These values were computed by excluding inapplicable cases.

$^b$ The values for $n = \infty$ were obtained theoretically.
Table 3
Results Obtained from Gojobori et al.'s (1982a) Computer Simulation

<table>
<thead>
<tr>
<th>TRUE δ AND n</th>
<th>TK METHOD</th>
<th>GIN METHOD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>δ</td>
<td>δₖ</td>
</tr>
<tr>
<td>1.0:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>144</td>
<td>1.00a</td>
<td>.24a</td>
</tr>
<tr>
<td>500</td>
<td>1.07</td>
<td>...</td>
</tr>
<tr>
<td>500</td>
<td>1.07</td>
<td>...</td>
</tr>
<tr>
<td>2.0:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>144</td>
<td>1.36a</td>
<td>.30a</td>
</tr>
<tr>
<td>500</td>
<td>1.53a</td>
<td>...</td>
</tr>
</tbody>
</table>

NOTE.—δ = average of the estimate (δ) of δ, δₖ = standard deviation of δ, f = proportion of inapplicable cases (denominator indicates the number of replications); n = number of nucleotide pairs.

* These values were computed by excluding inapplicable cases.

Table 4
Observed Numbers of the 10 Different Pairs of Nucleotides between the DNA Sequences for the Human and Rat Insulin A and B Chains

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AT</th>
<th>AG</th>
<th>AC</th>
<th>TT</th>
<th>TG</th>
<th>TC</th>
<th>GG</th>
<th>GC</th>
<th>CC</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>First position</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>13</td>
<td>51</td>
</tr>
<tr>
<td>Third position</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>8</td>
<td>2</td>
<td>21</td>
<td>51</td>
</tr>
</tbody>
</table>

NOTE.—The numbers at the first and third nucleotide positions of codons are listed separately. There are no nucleotide differences at the second position.

Consider that the number of replications is only 50, the agreement between the observed and expected values seems to be reasonably good. Table 3 also gives the observed standard deviations for the TK and GIN methods. They are relatively small compared with those for (3) and (6) because there were many inapplicable cases excluded.

Numerical Example

Sures et al. (1980) determined the nucleotide sequence of the human preproinsulin mRNA and compared it with that of the rat preproinsulin-I mRNA. Preproinsulin consists of four polypeptide chains—the A and B chains, signal peptide, and C peptide. The A and B chains (51 amino acids) produce active insulin, whereas the signal and C peptides (54 amino acids) are removed before insulin is produced. Since the latter two polypeptides are considered to be subject to less stringent purifying selection than the former two polypeptides (Sures et al. 1980), we have analyzed them separately. Following Kimura (1981), we have also considered the first, second, and third nucleotide positions of codons separately. The numbers of 10 different pairs of nucleotides (nᵢ) between the DNA sequences for the human and rat A and B chain genes are given in table 4. (The mRNA sequences were converted into the DNA sequences.) The relative frequency of nucleotide pair i and j (xᵢₗ) can then be obtained by dividing these numbers (nᵢₗ) by the total number, that is, 51. Once the xᵢₗ’s are obtained, the average frequency of the ith nucleotide for the two sequences under comparison (qᵢ) is given by qᵢ = xᵢₗ + Σᵢᵢⱼ xᵢⱼ/2. Thus, we obtain qₐ = 0.186, q₉ = 0.294, q₇ = 0.255, and q₉ = 0.265 for the first nucleotide position. We also have π = Σᵢ xᵢₗ.
(i < j) = 0.0392, b_i = 1 - \Sigma q_i^2 = 0.7437, h = 0.005978, b_2 = \pi^2/h = 0.2573, and 
\[ b = (b_1 + b_2)/2 = 0.5005. \] 
Thus, the estimate of \( \hat{\delta} \) is \( \hat{\delta} = 0.04 \) from (6). However, 
the variance of \( \hat{\delta} \) becomes 0.00087 from (9). Therefore, the standard error of \( \hat{\delta} \) is 
0.03. A similar computation for the third nucleotide position gives \( \hat{\delta} = 0.55 \pm 0.20 \). (There are no nucleotide differences at the second position.) It should be 
noted that in the present case application of the JC method gives \( \hat{\delta} = 0.04 \pm 0.03 \) for the first position and \( \hat{\delta} = 0.44 \pm 0.12 \) for the third position (table 5). 
Therefore, only when \( \hat{\delta} \) is sufficiently large does the difference between the two 
methods become appreciably large. The estimates obtained by the TK and GIN 
methods are also presented in table 5. These methods again give essentially the 
same result for the first position, but the estimates for the third position are larger 
than the estimate from (6).

Table 5 also includes the estimates of \( \hat{\delta} \) for the first, second, and third 
nucleotide positions for the signal and C peptides. At the first and second positions 
the four methods used all give essentially the same estimate of \( \hat{\delta} \). As expected, 
the \( \hat{\delta} \) values for the signal and C peptides are larger than those for the A and B 
chains. At the third position of the signal and C peptides the JC method gives \( \hat{\delta} = 0.63 \pm 0.16 \) and equation (6), \( \hat{\delta} = 0.91 \pm 0.39 \). The other two methods are 
not applicable to this case. The value of \( \hat{\delta} = 0.91 \) obtained by equation (6) is quite 
high compared with the corresponding value of the A and B chains. If we assume 
that the time since divergence between man and rat is \( 8 \times 10^9 \) years, this gives 
a rate of nucleotide substitution of \( 5.7 \times 10^{-9} \) per site per year. This is as high 
as Li et al.'s (1981) estimate (4.6 \( \times \) \( 10^{-9} \)) of the rate of nucleotide substitution 
for pseudogenes. It is possible that there is little purifying selection operating at 
the third positions for these peptides.

**Evolutionary Distance due to Deletion and Insertion**

Recent data on nucleotide sequences of related genes indicate that a sub-
stantial proportion of evolutionary change of DNA sequence arises from deletion 
and insertion of nucleotides, particularly in noncoding regions of DNA. We note 
that most deletions and insertions are short and occur with an appreciable fre-
quency (e.g., Efstratiadis et al. 1980; Langley et al. 1982; Cann and Wilson 1983). 
It is therefore possible to study the effects of these events on DNA divergence.

### Table 5

**Estimates (\( \hat{\delta} \)) of the Number of Nucleotide Substitutions per Site between the**

**Human Preproinsulin and Rat Preproinsulin I Genes at the First, Second,**

**and Third Nucleotide Positions of Codons**

<table>
<thead>
<tr>
<th>Gene Region and Position in Codon</th>
<th>JC Method</th>
<th>GIN Method</th>
<th>TK Method</th>
<th>Equation (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A + B chains (( n = 51 ))</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First</td>
<td>.04 ± .03</td>
<td>.04 ± .03</td>
<td>.04 ± .03</td>
<td>.04 + .03</td>
</tr>
<tr>
<td>Second</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Third</td>
<td>.44 ± .12</td>
<td>.60 ± .25</td>
<td>.79 + .53</td>
<td>.55 + .20</td>
</tr>
<tr>
<td>Signal + C peptides (( n = 54 ))</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First</td>
<td>.17 ± .06</td>
<td>.19 ± .08</td>
<td>.15 + .11</td>
<td>.18 ± .07</td>
</tr>
<tr>
<td>Second</td>
<td>.21 ± .07</td>
<td>.22 ± .08</td>
<td>.22 ± .07</td>
<td>.22 ± .08</td>
</tr>
<tr>
<td>Third</td>
<td>.63 ± 16</td>
<td>( \infty )</td>
<td>( \infty )</td>
<td>.91 ± .39</td>
</tr>
</tbody>
</table>

**Source.** — Data from Sures et al. (1980).

*\( \infty \) = inapplicable case.*
Nei et al. (1984) proposed a simple method of measuring the evolutionary distance between two homologous DNA sequences due to deletion and insertion: they compute the number of gap nucleotides per nucleotide site between a pair of DNA sequences compared. This quantity seems to be appropriate when a short period of evolutionary time is considered. When the evolutionary time considered is long, however, the following method seems to be better than that of Nei et al. (1984).

We again consider two homologous nucleotide sequences (X and Y) that diverged from a common ancestral sequence t evolutionary time units (e.g., years) ago. We assume that the length of a deletion or insertion is short compared with the total length of the DNA sequence (n) and that deletion and insertion occur independently. Let $\alpha$ be the proportion of DNA that is deleted during unit evolutionary time, i.e., $\alpha = m_p/n$, where $m_p$ is the number of nucleotides deleted and n is the total number of nucleotides before deletion. Note also that $\alpha$ is the number of nucleotide deletions per nucleotide site and usually a very small quantity. Similarly, we denote by $\beta$ the proportion of DNA that is inserted during unit evolutionary time, that is, $\beta = m_i/n$, where $m_i$ is the number of nucleotides inserted. We assume that $n$ remains more or less the same because of the compensating effects of deletion and insertion. In practice, $\alpha$ and $\beta$ may vary with evolutionary time, and we denote the values of $\alpha$ and $\beta$ for the $i$th evolutionary time unit by $\alpha_i$ and $\beta_i$, respectively. If we assume that deletion and insertion occur independently in sequences X and Y, the total number of nucleotide deletions and insertions per nucleotide site over the entire $t$ is given by

$$\gamma = \sum_{i=0}^{t-1}(\alpha_i + \beta_i)$$

$$= 2(\bar{\alpha} + \bar{\beta})t,$$

where $\bar{\alpha}$ and $\bar{\beta}$ are the averages of $\alpha_i$ and $\beta_i$, over evolutionary time, respectively. In this connection it should be noted that $\gamma$ measures only the DNA divergence due to deletion and insertion, and no consideration is given to the DNA changes due to nucleotide substitution.

The value of $\gamma$ can be estimated in the following way. We first consider the evolutionary change of the number of nucleotides (n) in the lineage of X. Let $n_X(t)$ be the total number of nucleotides at time $t$ in this lineage. We then have

$$n_X(t) = n_X(t-1)(1 - \alpha_i)(1 + \beta_i)$$

$$= n_X(0)\prod_{i=0}^{t-1}(1 - \alpha_i)(1 + \beta_i)$$

$$= n_X(0)e^{-2\alpha t + 2\beta t},$$

where $n_X(0)$ is the initial number of nucleotides. A similar expression can be obtained for n for Y, that is, $n_Y(t)$. However, the total number of homologous nucleotides shared by X and Y is given by

$$n_{XY}(t) = n_{XY}(t-1)(1 - \alpha_{i-1})^2$$

$$= n_X(0)e^{-2\alpha t},$$
because insertions do not create any homologous DNA segments. Therefore, we have

\[ P = \frac{n_{xy}}{\sqrt{n_x n_y}} = \exp \left( -\sum_{i=0}^{t-1} (\alpha_i + \beta_i) \right) \]  

(14)

where \( n_x, n_y, \) and \( n_{xy} \) are the observed values of \( n_x(t), n_y(t), \) and \( n_{xy}(t). \) Thus, \( \gamma \) in (11) can be estimated by

\[ \gamma = -2 \log P. \]  

(15)

It is noted that \( P \) can also be defined as

\[ P = 2n_{xy}(n_x + n_y). \]  

(16)

This definition is simpler than (14), but when the rates (\( \alpha \) and \( \beta \)) of deletion and insertion are not the same for sequences \( X \) and \( Y, \) (14) is more reasonable. In practice, however, (14) and (16) usually give very similar values.

Comparison with Nei et al.'s Formula

Nei et al. (1984) proposed to measure the DNA divergence due to deletion and insertion by

\[ \gamma_m = \frac{\bar{g}}{m_r}, \]  

(17)

where \( g \) is the number of nucleotides in the gaps between two DNA sequences and \( m_r \) is the total number of nucleotides compared. This gives a minimum estimate of DNA divergence due to deletion and insertion. This can be seen from figure 2, in which an artificial example of evolutionary change of DNA due to deletion and insertion is presented. In this example sequence \( X \) at time I has a deletion of 60 nucleotides (nt) starting from nucleotide position 301, whereas sequence \( Y \) has a deletion of 40 nt starting from position 601. Therefore, the divergence between \( X \) and \( Y \) is properly measured by \( \gamma_m, \) which becomes 100/1,000 = 0.1. In practice, however, we do not know the ancestral sequence of \( X \) and \( Y, \) so it is difficult to determine whether the two gaps between \( X \) and \( Y \) are due to deletion or insertion. If they are caused by insertion, the ancestral sequence should have had 900 nt instead of 1,000. In this case the DNA divergence should be 100/900 = 0.111. This indicates that \( \gamma_m \) gives an underestimate of DNA changes if both deletion and insertion occur. Our formula (15) takes care of both deletion and insertion, though it depends on the model used. In the present case the estimate (\( \hat{\gamma} \)) obtained by equation (15) is 0.108, which is intermediate between the two estimates obtained above.

Another advantage of \( \gamma \) over \( \hat{\gamma}_m \) is that it takes care of multiple events of deletion and insertion at least to some extent. In figure 2 sequence \( X \) experienced an insertion during the evolutionary period between time I and time II, whereas sequence \( Y \) experienced another deletion involving positions from 351 to 380. The latter deletion is overlapped with the deletion in \( X, \) so that \( \gamma_m \) gives an underestimate of DNA changes. It becomes 180/1,060 = 0.170. In (15) deletions and insertions are assumed to occur independently, and multiple deletions and insertions are taken into account. Indeed, \( \hat{\gamma} \) becomes \(-2 \log(880/\sqrt{1,010 \times 930}) = 0.193, \) which is larger than \( \gamma_m. \)
Numerical Example

Efstratiadis et al. (1980) compared the nucleotide sequences of various parts of the noncoding regions of globin genes from diverse organisms. This comparison indicates that a majority of deletions/insertions involve a small number of nucleotides, but there are a few deletions/insertions in which a large number of nucleotides (more than 50) are involved. However, amino acid sequence data suggest that deletions and insertions are much less frequent in the coding regions of globin genes than in the noncoding regions (Hunt et al. 1978). To see the pattern of accumulation of DNA changes due to deletion/insertion, we computed the evolutionary distance given by (15) for the 5' flanking region (including about 120 nt upstream starting from the cap site), 5' leader region (about 50 nt between the cap site and the initiation codon), intron I (about 130 nt), and 3' tail (noncoding) region (about 130 nt) of globin genes as well as for the coding region (about 438 nt or 146 codons). We used Efstratiadis et al.'s (1980) data for the noncoding region and Hunt et al.'s (1978) data for the coding region. In the latter case we used a codon rather than a nucleotide as a unit of change, because this does not change the numerical value of our measure. In both cases we assumed that the authors' alignment of sequences was correct.

The values of $n_x$, $n_y$, and $n_{xy}$ for the coding region (amino acid sequence) are presented in table 6. From these values we can estimate $\gamma$ by using (15). For example, in the case of human (X) and newt (Y) $\alpha$ chain genes $n_x = 141$, $n_y = 142$, and $n_{xy} = 141$. Therefore, $\gamma$ becomes 0.007. Table 6 indicates that $\gamma$ is small when the two sequences compared are closely related but tends to increase as the time since divergence ($t$) increases. Thus, the comparison of human and shark $\alpha$ chains gives a value of $\gamma = 0.084$. However, $\gamma$ does not seem to be linearly related to evolutionary time (fig. 3). Namely, $\gamma$ is 0 up to $t = 300$ million years (Myr) and then increases slightly. This reflects the fact that the length of the coding region of DNA is strongly conserved in the evolutionary process.

The noncoding region of DNA undergoes a much more rapid change due to deletion/insertion. However, the four different parts of the noncoding region seem

![Fig. 2.—A hypothetical example of evolutionary changes of DNA sequences due to deletion and insertion. Solid lines stand for DNA sequences, and broken lines, gaps. The numbers on DNA sequences represent nucleotide positions. See text for further explanation.](image-url)
Table 6
Estimates of Evolutionary Distances ($\gamma$) due to Deletion and Insertion among the Coding Region Sequences of Various Globin Genes (below the diagonal)

<table>
<thead>
<tr>
<th>Gene</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Human $\alpha$</td>
<td>141</td>
<td>141</td>
<td>140</td>
<td>139</td>
<td>139</td>
<td>139</td>
<td>134</td>
<td></td>
</tr>
<tr>
<td>2. Chicken $\alpha$</td>
<td>0</td>
<td>141</td>
<td>140</td>
<td>139</td>
<td>139</td>
<td>139</td>
<td>134</td>
<td></td>
</tr>
<tr>
<td>3. Newt $\alpha$</td>
<td>0.07</td>
<td>0.07</td>
<td>140</td>
<td>140</td>
<td>140</td>
<td>140</td>
<td>134</td>
<td></td>
</tr>
<tr>
<td>4. Carp $\alpha$</td>
<td>0.21</td>
<td>0.21</td>
<td>0.28</td>
<td>140</td>
<td>139</td>
<td>140</td>
<td>134</td>
<td></td>
</tr>
<tr>
<td>5. Shark $\alpha$</td>
<td>0.084</td>
<td>0.084</td>
<td>0.076</td>
<td>0.091</td>
<td>140</td>
<td>140</td>
<td>134</td>
<td></td>
</tr>
<tr>
<td>6. Human $\beta$</td>
<td>0.063</td>
<td>0.063</td>
<td>0.056</td>
<td>0.056</td>
<td>0.104</td>
<td>140</td>
<td>140</td>
<td></td>
</tr>
<tr>
<td>7. Chicken $\beta$</td>
<td>0.063</td>
<td>0.063</td>
<td>0.056</td>
<td>0.056</td>
<td>0.104</td>
<td>0</td>
<td>140</td>
<td></td>
</tr>
<tr>
<td>8. Frog $\beta$</td>
<td>0.095</td>
<td>0.095</td>
<td>0.102</td>
<td>0.102</td>
<td>0.150</td>
<td>0.042</td>
<td>0.042</td>
<td></td>
</tr>
</tbody>
</table>

NOTE.—In this table $\gamma$ represents the distance per codon rather than per nucleotide.
The values above the diagonal are the numbers of codons shared ($\alpha_{xy}$) by the two sequences compared (the total number of codons compared minus the number of codons in the gaps). The values on the diagonal are the number of codons in the sequence concerned ($\alpha_{xx}$ or $\alpha_{yy}$).

Fig. 3.—Relationships between the evolutionary distances ($\gamma$) for various parts of globin genes and evolutionary time. • = coding regions, ○ = 5' leader region, ▲ = intron I, △ = 5' flanking region, X = 3' tail (noncoding) region. To avoid overcrowding of data points, we present only the results for the comparisons involving human globins. $H\alpha$ = human $\alpha$ globin, $H\beta$ = human $\beta$ globin, $H\bar{\epsilon}$ = human $\epsilon$ globin, $M\beta_{mu}$ = mouse $\beta_{mu}$ globin, $H\gamma$ = human $\gamma$ globin ($\alpha_{\gamma}$ and $\alpha_{\gamma}$). $H\epsilon$ = human $\epsilon$ globin, $CH\alpha$ = chicken $\alpha$ globin, $N\alpha$ = newt $\alpha$ globin, $CA\alpha$ = carp $\alpha$ globin, and $S\alpha$ = shark $\alpha$ globin. The evolutionary times used are identical with those used by Efstratiadis et al. (1980) and Dayhoff (1972). MY = million years.

to have different rates of accumulation of DNA changes (fig. 3). The 3' tail region apparently has the highest rate, whereas the 5' leader region has the lowest rate. This is probably because the 5' leader region plays an important role for mRNA processing and translation and thus the DNA sequence is not very flexible. The relationship between $\gamma$ and evolutionary time is again nonlinear, though $\gamma$ generally increases as $t$ increases. This nonlinear relationship is mainly due to the fact that a deletion or insertion occasionally involves a large number of nucleotides. Thus,
the large value of $\gamma$ for the comparison of the 3' tail regions of the human $\epsilon$ and $\gamma$ chains is caused by the fact that the $\gamma$ chain has a long stretch of deletion (44 nt) compared with the $\epsilon$ chain.

**Discussion**

We have seen that our new formulas, particularly equation (6), give a good estimate of nucleotide substitutions as long as the true value of $\delta$ is less than 1. For $\delta > 1$, the GIN method seems to be better than equation (6), if we exclude the cases where the formulas are inapplicable. However, when $\delta > 1$, the GIN method is very often inapplicable because of a negative argument in the logarithm involved. Therefore, if we take into account this property as well as the simplicity of equation (6) compared with the GIN formula, (6) seems to be generally preferable to the GIN method. It should also be noted that in most studies of molecular evolution $\delta$ is smaller than 1, so that equation (6) can be applied to a wide variety of cases.

It should be noted, however, that equation (6) depends on the assumption that all nucleotide sites examined are subject to the same pattern of nucleotide substitution irrespective of the location of the nucleotide. In practice, this assumption does not seem to hold in many cases. It is well known that functionally important parts of genes are subject to nucleotide substitution less often than unimportant parts. Amino acid-altering nucleotide substitutions are also known to occur less frequently than synonymous substitutions. When the number of nucleotide substitutions per site ($\delta$) is small, this causes no problem, since there will be few backward and parallel substitutions in this case. As $\delta$ increases, however, backward and parallel substitutions may accumulate at functionally less important sites, whereas functionally more important sites may remain substitution free. In this case the method proposed here is expected to give underestimates of $\delta$. At the present time, it is not easy to take into account this factor properly, though some approximate treatment of the problem has been proposed (Nei and Li 1979). To make a general formulation of this problem, a more detailed knowledge of nucleotide substitution in various genes is required.

Our formulation of $\gamma$ in (15) was presented to quantify the effect of deletion and insertion on the evolutionary change of DNA sequences. As we have seen from data on globin genes, the evolutionary change of DNA arising from these factors occurs in a less regular fashion than that arising from nucleotide substitution. This is because there is a small proportion of large deletions and insertions that involves a large number of nucleotides. These deletions and insertions apparently occur haphazardly but affect the DNA sequences substantially once they occur. Because of this, $\gamma$ generally does not increase linearly with evolutionary time and thus cannot be used as a molecular clock. Nevertheless, $\gamma$ gives a quantitative measure of DNA change due to deletion and insertion and would be useful for evolutionary studies of DNA sequences.

**Acknowledgments**

We thank Dr. Clay Stephens for his comments on the manuscript. This work was supported by research grants from the National Institutes of Health and the National Science Foundation.
APPENDIX

Nucleotide Substitution under the Equal-Input Model

Let us denote nucleotides A, T, G, and C by 1, 2, 3, and 4, respectively. Let $\lambda_{ij}$ be the rate of substitution of the $j$th nucleotide for the $i$th nucleotide per unit evolutionary time (e.g., year) and $q_i$ be the equilibrium frequency of the $i$th nucleotide. In the equal-input model (Tajima and Nei 1982), $\lambda_{ij} = \lambda_{ji} = \lambda_{ij}^* = \lambda_j$, is assumed for all $\lambda_{ij}$ except $\lambda_{ii}$, which is equal to $1 - \Sigma_{i=1}^{4} a_i$ for $i \neq j$. Therefore, the transition matrix for the four nucleotides may be written as

$$
P = \begin{bmatrix}
1 - (a_2 + a_3 + a_4) & a_2 & a_3 & a_4 \\
a_1 & 1 - (a_1 + a_3 + a_4) & a_3 & a_4 \\
a_1 & a_2 & 1 - (a_1 + a_2 + a_4) & a_4 \\
a_1 & a_2 & a_3 & 1 - (a_1 + a_2 + a_3)
\end{bmatrix}, \quad (A1)
$$

and the equilibrium frequency of the $i$th nucleotide is given by

$$q_i = a_i / \sum_{j=1}^{4} a_j \quad . \quad (A2)$$

(Tajima and Nei 1982).

Let us now consider two long homologous nucleotide sequences ($X$ and $Y$) that diverged from a common ancestral sequence $t$ years (or evolutionary time units) ago. We denote by $y_{ij}(t)$ the proportion of homologous nucleotide pairs where $X$ and $Y$ have nucleotides $i$ and $j$, respectively, at time $t$. Then we have

$$y_{ij}(t) = \sum_{m=1}^{4} \sum_{n=1}^{4} \lambda_{mn} y_{mn}(t - 1). \quad (A3)$$

Under the equal-input model (A3) is approximately given by

$$y_{ij}(t) = \left(1 - \Sigma_{k \neq i} a_k - \Sigma_{k \neq j} a_k\right) y_{ij}(t - 1) + a_i \sum_{n \neq j} y_{in}(t - 1) + a_j \sum_{m \neq i} y_{jm}(t - 1). \quad (A4)$$

Using (A2), we obtain

$$y_{ij}(t) = \left[y_{ij}(0) - q_i q_j\right] \left(1 - 2 \Sigma_{k=1}^{4} a_k\right)^t + q_i q_j \quad . \quad (A5)$$

First consider the case of $i \neq j$. In this case $y_{ij}(0) = 0$, because at time 0 the two sequences must have been the same. Therefore, we have

$$y_{ij}(t) = q_i q_j \left[1 - \exp\left(-2 \Sigma_{k=1}^{4} a_k t\right)\right]. \quad (A6)$$

When $i = j$, we have $y_{ii}(0) = q_i$ and
\[ y_{ij}(t) = q_i(1 - q_j) \exp \left( -2 \sum_{k=1}^{4} a_{ik} t \right) + q_j^2. \]  

(A7)

Let us denote by \( x_{ij} \) the proportion of pairs of nucleotides \( i \) and \( j \) \( (i < j) \) between sequences \( X \) and \( Y \). When \( i \neq j \) (A6) gives

\[ x_{ij} = y_{ij}(t) + y_{ji}(t) = 2q_i q_j \left[ 1 - \exp \left( -2 \sum_{k=1}^{4} a_{ik} t \right) \right]. \]  

(A8)

This equation indicates that \( x_{ij}/(2q_i q_j) \) is constant for all combinations of \( i \) and \( j \) \( (i < j) \).

The average number of nucleotide substitutions per site between sequences \( X \) and \( Y \) is

\[ \delta = 2 \sum_{i=1}^{4} q_i (1 - \lambda_i) t. \]  

(A9)

Under the equal-input model it becomes

\[ \delta = 2 \left( 1 - \sum_{i=1}^{4} q_i^2 \right) \sum_{i=1}^{4} a_{ii} t. \]  

(A10)

Substitution of (A10) into (A8) gives

\[ x_{ij} = 2q_i q_j \left\{ 1 - \exp \left[ -\delta \left( 1 - \sum_{i=1}^{4} q_i^2 \right) \right] \right\}. \]  

(A11)

Since \( \pi = \sum_{i} x_{ij} \) and \( b_i = 2\sum_{j<i} q_i q_j = 1 - \sum_{i} q_i^2 \) for \( i < j \), we obtain (3) in the text. From (A11) we also have

\[ \delta = - \left( 1 - \sum_{i=1}^{4} q_i^2 \right) \log_e [1 - x_{ij}/(2q_i q_j)]. \]

Since \( x_{ij}/(2q_i q_j) \) is constant for all values of \( i < j \), we obtain (4) in the text.

LITERATURE CITED


Walter M. Fitch, reviewing editor

Received August 29, 1983; revision received October 17, 1983.
Book Reviews


Recent advances in molecular biology have revolutionized our understanding of eucaryotic gene and genome structure, and revealed novel mechanisms of evolution at the molecular level. This has been the result of the rapid accumulation of DNA primary sequences and restriction maps. To population geneticists and molecular evolutionists, statistical analyses of these data are crucial, and the nine chapters in this book are a survey of these statistical methods.

Chapter 1, by Schaffer, deals with regression methods that predict DNA fragment lengths from mobilities on a gel. This problem is not difficult to solve statistically but it does have some practical importance. Gingeras, in chapter 2, discusses various computer algorithms for DNA sequence analysis but fails to tell us where the statistical problems do occur. He provides a useful list of available computer programs, but there is no discussion of the merits of each. However, he does mention the possibility of using artificial intelligence techniques for sequence analysis, and I, for one, would have liked to see a detailed discussion of this interesting area.

The next two chapters, by Ewens and Kaplan, respectively, are reviews of models and methods for studying evolutionary relatedness and genetic variability in populations based on restriction maps or DNA sequences. Ewens goes to great length to point out the several types of variability that one could consider and when each is appropriate; I found this discussion very useful. However, Kaplan does a more comprehensive job of studying the various estimators by computer simulations and data analysis.

Chapter 5, by Brown and Clegg, on the analysis of variation in related sequences, is perhaps the best in this collection. These authors provide new statistical methods and perform a comprehensive analysis of data on a repeated DNA sequence in maize knob heterochromatin. They repeatedly point out where the statistical problems lie, what they are, and what approaches may be taken to solve them. I was particularly interested in their method of inferring concerted evolution. In the next chapter, Felsenstein, a leader in the field of phylogenetic analysis, has described the statistical problems in this area very well. These two chapters are a must for statisticians interested in a new class of problems and are areas where they have much to contribute.

Errors in phylogenetic analysis that can occur through chance convergent evolution are studied by Templeton in chapter 7, and he suggests nonparametric methods as a remedy. I disagree with much of what he has to say. First, his probabilities are all conditional on the type of convergent evolution—a fact we will not know. Second, the simulation studies of Kaplan (chap. 4) demonstrate that current methods work rather well, and it behooves Templeton to show the superiority of his methods by computer simulation.

Chapters 8 and 9 contain discussions of the use of DNA data in human genetics. Bishop et al. study the number of DNA polymorphisms necessary to map the human genome so that an unknown disease locus can be shown to be linked to such a marker. These studies are interesting, but it is not clear whether gene mapping will proceed using the strategy that these authors recommend. The last chapter, by Asmussen and Clegg, discusses the use of DNA markers in
prenatal diagnosis of genetic diseases. The results are correct, but I found their evolutionary analysis of the usefulness of marker genes irrelevant in the context of medical genetics.

This book has been printed attractively and contains few typographical errors. However, on page 53, formula 18 should be $\theta/((1 + \theta)\log n)$, and on page 97 the $O(\theta)$ should be $O(n)$. The editor has done an excellent job but he has slipped once—the reference to a human-rodent hybrid (line 9, p. 183)! I presume this is a human-rodent cell hybrid.

This is the first book that has attempted to gather statistical methods for analysis of DNA data, and I enjoyed it, but I also expected a great deal more. Considerable DNA sequence data exist, and it would have been useful to include more analyses than provided. This book will be useful to students of genetics and evolution, particularly for newcomers.

ARAVINDA CHAKRAVARTI
University of Pittsburgh


The early ancestors of all DNA-containing organisms had only a few genes. Today, the amount of DNA per genome encompasses four orders of magnitude from about $4 \times 10^6$ nucleotide pairs (np) in bacteria and $4 \times 10^9$ in fungi to more than $10^{10}$ in some primitive fishes, salamanders, and many plants. Polyploidy is one process by which genome size increases. Wen-Hsiung Li reviews other, more fundamental mechanisms that have been elucidated in recent years. One mechanism is gene elongation, an increase in gene size that may occur by tandem duplication of relatively short nucleotide sequences. An extreme example is a gene coding for collagen in chickens that consists of more than 50 exons each made up of five to a dozen tandem duplications of one basic sequence 9 np long. Complex genes evolve as well from simpler ones by the joining of small primordial genes having separate functions. The ancestral genes may be recognized in the different exons, each coding in the modern gene for a distinct protein domain, as in the gene encoding the constant region of the heavy chain of immunoglobulin $\gamma$.

Complete genes also become duplicated in evolution. Some, like those coding for the ribosomal or transfer RNAs, exist in multiple copies that have remained identical with one another in structure and function. Others, such as the globin genes in vertebrates, diverge after duplication and acquire novel but related functions. And then, there are nucleotide sequences of unknown function that are multiplied many times in a genome; for example, the Alu sequences, of which there are about 300,000 in the human genome, each some 300 np long. The evolution of the DNA by duplication is a fascinating problem. I guess that we know only the tip of the iceberg and that exciting discoveries will be forthcoming.

One puzzle is that some duplicated genes or nucleotide sequences remain identical for eons. This phenomenon is now called concerted evolution and may occur by gene conversion or unequal crossing-over. Norman Arnheim argues convincingly against a possible alternative to these processes, namely, natural selection acting separately but in parallel in each of the duplicated sequences. Particularly difficult to understand is the conservation of homology in sequences located in nonhomologous chromosomes, a problem explored by Arnheim using the human ribosomal genes as a model.

Classical genetics established that genes generally occupy fixed locations on chromosomes. The 1983 Nobel Prize in physiology or medicine was awarded to Barbara McClintock for demonstrating that some genetic elements move from
Allan Campbell distinguishes *episomes*, which can replicate independently of the chromosome, from *transposons*, which are found only in the inserted state and are of two types: true transposons, which include genes that determine phenotypic traits; and *insertion sequences*, which have no other known properties besides transposability and promoter or terminator activity. Transposons affect chromosome evolution because they induce rearrangements such as deletions and inversions. Modification of gene regulation may very well be the most lasting effect of transposons on evolution.

Restriction endonuclease digestion and DNA sequencing have laid open the exploration of organelle DNA. The large-scale evolution of animal mitochondrial DNA is reviewed by Wesley M. Brown, who argues for the stability of gene order for hundreds of millions of years, albeit differences exist between insects and vertebrates. These two large groups of organisms differ in DNA composition as well: G and C content amounts to somewhat more than 40% of the mitochondrial DNA of most vertebrates but to only 21% in *Drosophila melanogaster*. Nevertheless, as reviewed by John C. Avise and Robert A. Lansman, the mitochondrial DNA of animals exhibits high-sequence polymorphism within species—several times greater than for nuclear DNA. This mitochondrial polymorphism is mostly due to base substitutions rather than additions or deletions; and transitions prevail over transversions. The situation is altogether different in plants, where structural reorganization of the mitochondrial DNA appears to be common, but sequence homology is largely preserved.

Genetic polymorphism within species is further examined in other chapters. Robert K. Selander and Thomas S. Whittam review the most recent results concerning electrophoretically cryptic protein variation and present a valuable discussion of what we have learned about population structure from the distribution of protein polymorphisms in snails and in humans. Richard K. Koehn, Anthony J. Zera, and John G. Hall examine the contribution of natural selection to the maintenance of polymorphisms, whereas Masatoshi Nei emphasizes the significance of mutation from a neutralist point of view. Motoo Kimura restates with conviction his neutrality theory of molecular evolution in the light of the recent accumulation of DNA sequences.

*Evolution of Genes and Proteins* is a gem. It derives from a symposium held in June 1982 at the State University of New York in Stony Brook. Following strict editing rules, Nei and Koehn have gone as far as seems reasonable in order to integrate the separate contributions of many authors into a coherent book. The book, to the credit of editors and publisher, is well produced. Now may be high time to acknowledge Sinauer as an intelligent and innovative publisher that is contributing quite significantly to the current evolutionary literature.

Francisco J. Ayala University of California, Davis


This volume, the first in (another) new series of costly monographs, is based on a symposium held in 1980 at the Third International Congress of Systematic and Evolutionary Biology convened in Vancouver, British Columbia. It consists of nine chapters divided into three major sections: I, Eutherian phylogeny and protein evolution (chaps. 1–5); II, Modeling the process of sequence divergence (chaps. 6, 7); and III, Prospects for investigating evolution through genomic DNA (chaps. 8, 9). Chapters 1, 8, and 9 were not part of the Vancouver symposium but were added subsequently to provide nonmolecular information on eutherian
phylogeny (chap. 1) and to describe the newest methods of dissecting and reconstructing genomic DNA evolution (chaps. 8, 9).

The five chapters constituting part I account for nearly 60% of the book. M. Novacek reviews background anatomical and fossil evidence bearing on higher eutherian phylogeny. Five major areas of instability in eutherian phylogeny are identified and discussed in light of evidence from comparative anatomy and paleontology: (a) the Edentata, (b) the Carnivora, (c) the possible monophyly of the Rodentia, Lagomorpha, and related taxa, (d) the validity of the Archonta, and (e) the ungulate radiation. A composite cladogram for the major eutherian groups is presented and discussed. The chapter ends with a discussion of estimated divergence times for the eutherian orders and estimated rates of evolution among these taxa concluding "both molecular and taxonomic assessments of evolutionary rates in eutherians are limited critically by the imprecision of estimates of divergence dates from fossil and phylogenetic data" (p. 32).

J. J. Beintema and J. A. Lenstra discuss the "Evolution of mammalian pancreatic ribonucleases" based on amino acid sequence information from 35 mammalian species. Phylogenetic trees for the enzymes as well as trees obtained by using other biological information are presented, compared, and discussed—all these trees differ only slightly. Rates of evolutionary change in the ribonucleases are estimated, and considerable variation in evolutionary rates is noted for different taxa. Some discussion is also given to the three-dimensional structure and enzyme activities of the ribonucleases.

W. W. De Jong contributes a summary of comparative studies of vertebrate eye lens proteins. Particular attention is paid to the alpha-crystallin chains, and a summary of amino acid sequence comparisons of lens proteins of 41 mammalian species is presented. The goal is to "establish the most probable pattern of relationships among the mammalian orders" (p. 96). De Jong found that many questions could not be resolved because the proteins studied had evolutionary rates too slow to provide sufficient data to address questions of branching order among the taxa of interest. This slow evolutionary rate is acknowledged as responsible for the highly irregular rate of substitution in the alpha-crystallin proteins.

The longest chapter, by M. Goodman, A. E. Romero-Herrera, H. Dene, J. Czelusniak, and R. E. Tashian, summarizes the contributions amino acid sequence information has made to our understanding of mammalian, especially primate, phylogeny. The logic and methodology of the maximum parsimony algorithm used by Goodman’s laboratory for reconstructing phylogenies are reviewed. Both gene lineages and species lineages are reconstructed. Data from "553 polypeptide chains of 244 species" are analyzed and discussed. The recurrent theme throughout this analysis is the intrinsic variance observed in molecular evolution. It is clear from the data that some of the molecules studied exhibit a greater variance in their rates of evolution than others. Nevertheless, it is impressive that a correlation of .88 is obtained between the divergence times of mammalian lineages estimated from the fossil data, and from the overall molecular clock estimates (table 4). (Viewing molecular evolution as a statistical process, with a mean and a variance, seems to be a more powerful concept than viewing it as an absolute clock that sometimes runs fast and sometimes runs slowly.) Goodman et al.'s parsimony analysis identifies two periods of very rapid protein evolution which they suggest correlate with (1) the emergence of the tetrapods and (2) the origins of primates and the earliest eutherians. Much of this discussion has been presented before (references in bibliography), but it is all summarized here and will be of interest to new students of molecular evolution.

L. T. Hunt and the late M. O. Dayhoff cover "Evolution of chromosomal proteins" in the concluding chapter of section I. A brief review of eukaryote
chromosome structure and composition is presented, followed by a description of how computers can be used to detect homologous sequences within and between proteins. These data are then used to investigate phylogenetic relationships among proteins and protein families. The majority of this chapter focuses on the evolution of the histones, with evolutionary trees for histones H1-H5, as well as a phylogeny that includes all of the bovine nucleosome core histones. Trees are also described based on other chromosomal proteins including protamines and nonhistone chromosomal proteins.

Part II consists of a computer simulation study of the evolution of five proteins (M. Coates and S. Stone) and a study of gene and m-RNA structure in three protein families and attempts to define constraints on gene structure to be incorporated into their nonrandom REH evolutionary theory (R. Holmquist, D. Pearl, and T. H. Jukes). Both papers are concerned with modeling the process of molecular evolution to continue assessing whether significant amounts of evolution are indeed “neutral.” A major conclusion of Coates and Stone’s study is that each protein must be independently tested for its “clocklike” behavior—something that this reviewer had believed was common knowledge for the past decade! Perhaps a reason for continuing “molecular controversies” is the presentation of assertions that no molecule can be used as a clock because some molecules have an unacceptably high variance in their rates of molecular evolution.

The final section of the book was not part of the 1980 symposium but was added to present new methodologies available for the study of gene evolution. A. F. Scott and K. D. Smith describe restriction enzyme mapping of DNA as well as methods of studying all classes of repetitive DNA and single-copy DNA by hybridization analysis. A brief discussion of transposable elements and their role in evolution is also presented. In the concluding chapter D. Hewett-Emmett, P. J. Venta, and R. E. Tashian lucidly describe recombinant DNA technologies from cloning through nucleic acid sequencing and analysis. The evolutionary roles of gene structure (introns, exons, pseudogenes) and gene duplication are also ably presented.

Overall, this book is well presented, with few noticeable errors or typos. However, it is one of an ever-increasing proliferation of expensive volumes of symposia held years before the final publication date. Consequently, much of the information in this volume has already appeared in the reviewed literature. In contrast, other symposia dealing with molecular evolution have been published more rapidly and in paperback at much reduced prices. Such an approach to disseminating symposia papers seems highly preferable in being more timely and more readily available to graduate students and faculty alike.

LINDA R. MAXSON
University of Illinois at Urbana-Champaign
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 1 1/2 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.
Information for Contributors

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

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

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

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

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

Where the alignments disagree, they are differences rather than changes since there may have been multiple changes to create a single difference. Differences or changes are replacements if the sequences are amino acids, substitutions if they are nucleotides. Mutations should be restricted to changes before selection has operated. Homology must be defined since it has two common meanings: (1) observed similarity and (2) inferred common ancestry. The term similarity is preferred for meaning 1 because sequences may have similarity acquired by convergence (analogy) rather than retained after divergence (homology). When homology arises via a gene duplication (all or part), it is properly called paralogy; when it arises via speciation, it is properly called orthology. Gaps are introduced into sequences to increase their similarity rather than to optimize similarity (homology), unless an algorithm is employed that guarantees an optimized result according to the way similarity (homology) is defined (e.g., as maximum matches—a third meaning of homology). Similarity should not be asserted to be significant unless patently obvious or accompanied by a probability statement and its method of determination (χ², 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.
Information for Contributors

(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:


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 4½ inches wide by 7½ inches long (122 by 194 mm) in order to fit the journal format. Please keep in mind that these dimensions are maxima. Because of the need for a figure legend, the illustration cannot be the maximum size in both dimensions.

Proofs and Reprints

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

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

Managing Editor
Masatoshi Nei
Center for Demographic and Population Genetics
University of Texas Health Science Center at Houston
P.O. Box 20334
Astrodome Station
Houston, Texas 77025

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

Wesley M. Brown
Department of Cellular and Molecular Biology
University of Michigan
Ann Arbor, Michigan 48109

Richard B. Flavell
Department of Cytogenetics Plant Breeding Institute
University of Cambridge
Cambridge C82, 2LQ England

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

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

Robert K. Selander
Department of Biology
University of Rochester
Rochester, New York 14627
Further Insight into the Never-Ending Controversy over the Theory of Evolution...

Paperbacks From Washington Square Press

**THE BOOK OF DARWIN**
George Gaylord Simpson, editor
43126-9/$6.95

With an introduction and commentary by the Editor. The first comprehensive collection from the life and work of Darwin. "Dr. Simpson’s incomparable personal experience in evolutionary studies has produced a most valuable contribution to the works by and on Charles Darwin. . ."—Richard E. Leakey, author of *Origins* and *The Making of Mankind*

**THE MONKEY BUSINESS: A Scientist Looks at Creationism**
Niles Eldredge
44115-9/$2.95

The first book to explain the aims and methods of the movement called scientific creationism, by a paleontologist and curator at New York City’s Museum of Natural History. "...Eldredge marshals extensive evidence—from scientific findings to philosophical analyses—to demonstrate that ‘creation-science isn’t science at all, nor have creation scientists managed to come up with even a single intellectually compelling scientifically testable statement about the natural world.’”—The Washington Post Book World

---

**Costa Rican Natural History**
Edited by Daniel H. Janzen

In this long-awaited volume, 174 distinguished biologists explore the rich biological diversity of Costa Rica. Included are more than 300 species accounts plus substantial introductions, species checklists, and keys.

"This is an extraordinary, virtually unique work that will make an important contribution to our understanding of tropical biology not only in Costa Rica but throughout the world. The tremendous amount of original, previously unpublished, firsthand information is remarkable."—Peter H. Raven, Missouri Botanical Garden

Cloth $50.00  Paper $30.00  832 pages  306 photographs, 54 line drawings

The University of Chicago Press  Chicago, IL 60637
If the key word is genetics—

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

—the key research is in

The American Journal of Human Genetics

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

Editor: Dr. David E. Comings.

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

Subscribe to the AJHG.
In 1867, four renegades founded a journal.

Of the four men who left the Museum of Comparative Zoology in Cambridge to found The American Naturalist, two also participated in the founding, sixteen years later, of the American Society of Naturalists. The Society and the journal have worked closely together ever since, furthering the advancement and diffusion of knowledge about organic evolution and other broad biological principles.

In 1983, AN is still at the forefront.

One of the oldest scientific periodicals in the United States, The American Naturalist is also one of the foremost publications in the world for research in ecology, evolution, and population biology. Drawing upon examinations of all biotic kingdoms, the Naturalist emphasizes sophisticated analyses, innovative theoretical syntheses, and empirical discoveries with broad general implications.

A round table on research in ecology and evolutionary biology


Published monthly. George W. Salt, Editor.

The American Naturalist
"The book has no competition whatsoever. There is no anthology which even attempts to cover this ground."

—Richard Boyd, Cornell University

New in paperback

**Conceptual Issues in Evolutionary Biology**
edited by Elliott Sober

This collection of readings is designed for classroom use in the growing field of evolutionary biology. It assembles articles by outstanding philosophers and biologists who seek to clarify and connect some of the fields major concepts.

Articles are grouped into sections covering guiding ideas in evolutionary biology, fitness, units of selection, adaptation, function and teleology, the reduction of Mendelian genetics to molecular biology, and the nature of species.


A Bradford Book.

668 pp. $19.95 (cloth, $40.00)

New

**Genes, Organisms, Populations**
Controversies over the Units of Selection
edited by Robert N. Brandon and Richard M. Burian

This anthology collects some of the most important papers on natural selection. These are divided into three main sections: Historical Readings (Michael Ruse, Sewell Wright, V. C. Wynne Edwards, George C. Williams, Ernst Mayr); Levels and Units of Selection: Conceptual Analyses (William Wimsatt, Elliott Sober, Richard Lewontin, Robert N. Brandon, David Hull, Richard Dawkins; Models of Selection: Kin, Group, and Hierarchical (W. D. Hamilton, Richard E. Michod, J. Maynard Smith, Michael Wade, D. S. Wilson, Anthony J. Arnold, Jurt Fristrup)

A Bradford Book.

300 pp. $29.95

Now available in paperback

**Abusing Science**
The Case Against Creationism
Philip Kitcher

"Anyone who sincerely wants to understand evolutionary theory and the nature of science will find both explained simply and clearly in Kitcher's book. . . . Kitcher wrote his book as a 'manual for intellectual self-defense,' to be used by anyone called upon to respond to the bright-eyed onslaught of Creationists. It should also be read by anyone who thinks that the American sense of fair play requires that Creation Science be given equal time in biology classrooms." —The Quarterly Review of Biology

$6.95

Write for brochure

28 Carleton Street
Cambridge, MA 02142
Forthcoming

Immunological Similarities between Specific Chloroplast Ribosomal Proteins from Chlamydomonas reinhardtii and Ribosomal Proteins from Escherichia coli
Robert J. Schmidt, John E. Boynton, Nicholas W. Gillham, and Alan M. Myers

Molecular Probes of Phylogeny and Biogeography in Toads of the Widespread Genus Bufo
Linda R. Maxson

Molecular Evolution of Chloroplast DNA Sequences
Stephanie E. Curtis and Michael T. Clegg

Evolution of Glucagon Genes
Linda C. Lopez, Wen-Hsiung Li, Marsha L. Frazier, Chi-Cheng Luo, and Grady F. Saunders

Rabbit Globin Pseudogene \( \psi \beta 2 \) Is a Hybrid of \( \delta \)- and \( \beta \)-globin Gene Sequences
Ross C. Hardison and Jean B. Margot

Comparison of the \( \beta \)-like Globin Gene Families of Rabbits and Humans Indicates that the Gene Cluster 5'-\( \varepsilon \)-\( \gamma \)-\( \delta \)-\( \beta \)-3' Predates the Mammalian Radiation
Ross C. Hardison

Comparative Analysis of Human and Bovine Papillomaviruses
Samuel Karlin, Ghassan Ghandour, David E. Foulser, and Laurence J. Korn

The Bacterial Transposon Tn5: Evolutionary Inferences
Douglas E. Berg, Claire M. Berg, and Chihiro Sasakawa
Molecular Biology and Evolution

Editor in Chief
Walter M. Fitch, University of Wisconsin

Managing Editor
Masatoshi Nei, University of Texas

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

Editorial Board
Fred W. Allendorf
University of Montana
Norman Arnheim
SUNY at Stony Brook
John C. Avise
University of Georgia
Francisco J. Ayala
University of California, Davis
Anthony Hugh D. Brown
CSIRO, Canberra
L. L. Cavalli-Sforza
Stanford University
Elizabeth A. Craig
University of Wisconsin—Madison
Irving P. Crawford
University of Iowa
W. Ford Doolittle
Dalhousie University
Gabriel Dover
Cambridge University
Marshall H. Edgell
University of North Carolina
George Fox
University of Houston
Morris Goodman
Wayne State University
R. Grantham
Université de Lyon
Michael Grunstein
University of California, Los Angeles
Barry G. Hall
University of Connecticut
Daniel L. Hartl
Washington University
Alec Jeffreys
University of Leicester
Motoo Kimura
National Institute of Genetics, Japan
Costas Krimbas
Agricultural College of Athens, Greece
Charles H. Langley
National Institute of Environmental Health Sciences, North Carolina
Richard C. Lewontin
Harvard University
Wen-Hsiung Li
University of Texas at Houston
Alan Maxam
Harvard Medical School
Roger Milkman
University of Iowa
Takashi Miyata
Kyushu University, Japan
Tomoko Ohta
National Institute of Genetics, Japan
Dennis A. Powers
Johns Hopkins University
Nils Ryman
University of Stockholm, Sweden
Barbara Schaal
Washington University
Temple F. Smith
Northern Michigan University
Howard M. Temin
University of Wisconsin—Madison
Michael S. Waterman
University of Southern California
Sherman Weissman
Yale University Medical School
Gregory S. Whitt
University of Illinois at Urbana-Champaign
Eleutherios Zouros
Dalhousie University
Molecular Biology and Evolution
Volume 1, Number 3, April 1984

221 Individual and Evolutionary Variation of Primate Ribosomal DNA Transcription Initiation Regions
Golder N. Wilson, Mechthilde Knoller, L. Lynne Szura, and Roy D. Schmickel

238 Silencing of Duplicate Genes: A Null Allele Polymorphism for Lactate Dehydrogenase in Brown Trout (Salmo trutta)
Fred W. Allendorf, Gunnar Ståhl, and Nils Ryman

249 Calibrating the Molecular Clock: Estimates of Ground Squirrel Divergence Made Using Fossil and Geological Time Markers
David Glenn Smith and Richard G. Coss

260 The Relationship between Codon Boundaries and Multiple Reading-Frame Preferences: Coding Organization of Bacterial Insertion Sequences
David J. Galas and Temple F. Smith

269 Estimation of Evolutionary Distance between Nucleotide Sequences
Fumio Tajima and Masatoshi Nei

286 Book Reviews
Statistical Analysis of DNA Sequence Data, edited by B. W. Weir
Aravinda Chakravarti

Evolution of Genes and Proteins, edited by Masatoshi Nei and Richard K. Koehn
Francisco J. Ayala

Macromolecular Sequences in Systematic and Evolutionary Biology, edited by Morris Goodman
Linda R. Maxson