Review Article

Slipped-Strand Mispairing: A Major Mechanism for DNA Sequence Evolution

Gene Levinson and George A. Gutman
Department of Microbiology and Molecular Genetics, University of California, Irvine

Simple repetitive DNA sequences are a widespread and abundant feature of genomic DNA. The following several features characterize such sequences: (1) they typically consist of a variety of repeated motifs of 1-10 bases—but may include much larger repeats as well; (2) larger repeat units often include shorter ones within them; (3) long polypyrimidine and poly-CA tracts are often found; and (4) tandem arrangements of closely related motifs are often found. We propose that slipped-strand mispairing events, in concert with unequal crossing-over, can readily account for all of these features. The frequent occurrence of long tandem repeats of particular motifs (polypyrimidine and poly-CA tracts) appears to result from nonrandom patterns of nucleotide substitution. We argue that the intrahelical process of slipped-strand mispairing is much more likely to be the major factor in the initial expansion of short repeated motifs and that, after initial expansion, simple tandem repeats may be predisposed to further expansion by unequal crossing-over or other interhelical events because of their propensity to mispair. Evidence is presented that single-base repeats (the shortest possible motifs) are represented by longer runs in mammalian introns than would be expected on a random basis, supporting the idea that SSM may be a ubiquitous force in the evolution of the eukaryotic genome. Simple repetitive sequences may therefore represent a natural ground state of DNA unselected for coding functions.

Introduction

With the rapid accumulation of DNA sequence data in recent years, it has become apparent that a wide variety of simple repetitive motifs are commonly found in eukaryotic DNA. Both short and long tracts of simple repetitive DNA (SR-DNA) occur frequently at a variety of chromosomal loci within a broad range of organisms; the longer tracts are found mostly in higher eukaryotes. Highly repetitive tracts of considerable length have also been found in the genome of the yeast Saccharomyces (Bloom et al. 1982; Nakaseko et al. 1986; Wildeman and Nazar 1986). Hybridization studies have shown that SR-DNA is ubiquitous in a variety of genomes (Tautz and Renz 1984a, 1984b). The simple repeat poly-CA, for example, has been found in 70% of the clones of a mouse genomic library (Jeang and Hayward 1983) and is frequently found in many other genomic contexts as well (Hamada et al. 1982a, 1982b; Rogers 1987).

1. Key words: tandem duplications, tandem repeats, palindromes, simple DNA, repetitive DNA, satellite DNA, unequal crossing-over, recombination, insertions, deletions, frameshifts, mutations, control of gene expression. Abbreviations: [GATA]_2 and [GACA]_2 refer to tandem repeats of GATA and GACA, respectively; these and other sequences for which only one strand is shown always include, by implication, their complementary strand of the DNA double helix.

Address for correspondence and reprints: Dr. Gene Levinson, Department of Microbiology and Molecular Genetics, University of California, Irvine, California 92717.
FIG. 1.—Naturally occurring simple repetitive sequences. All sequences represent duplex DNA, but only one strand is shown. Included are examples of runs of a single base (A); tandem reiterations with repeats of ≥2 bases (B, C, and D); imperfect (quasi-) reiterated sequences (D, E); and sequences with homopolymers.
1983; Schmid and Shen 1985). Polypyrimidine tracts (polypurines on the complementary strand) are also abundant (see, e.g., Straus and Birnboim 1976; Dodd and Straus 1982; Heilig et al. 1982; Sorge and Hughes 1982; Schmid and Shen 1985). Computer analysis of sequence data has also demonstrated the prevalence of simple tandem repeats (Blaisdell 1983; Tautz et al. 1986; G. Levinson and G. A. Gutman, unpublished data), which are sometimes cryptic because of imperfections in the repeat units. Some examples of simple repetitive DNA tracts found in various published sequences are displayed in figure 1. (For more examples, see Slightom et al. 1980; Spritz et al. 1980; Dodd and Straus 1982; Hamada et al. 1982a, 1982b; Miklos and Gill 1982; Rodakis and Kafatos 1982; Singer 1982; Sorge and Hughes 1982; Moore 1983; Cato et al. 1984; Hasson et al. 1984; Rodakis et al. 1984; Skowronska et al. 1984; Miklos 1985; Willard et al. 1985; Chapman et al. 1986; Nakaseko et al. 1986; Wildeman and Nazar 1986).

Although the identification of SR-DNA has been widely reported, its origin and significance remain a mystery. One of the more puzzling features of these sequences is their diversity: tracts of SR-DNA may differ considerably in their organization, length, and base composition. However, a variety of simple motifs (such as the poly-CA and polypyrimidine tracts mentioned above) seem to occur repeatedly in SR-DNA in diverse contexts. In a previous study, we have shown that the simple repeats [GATA]n and [GACA]n—found in the genomes of taxa as distant as flies (Drosophila), snakes (Bungarus and Elaphe), and mice (Mus)—most likely evolved independently, possibly by a mechanism involving slipped-strand mispairing (SSM) of the two strands of the DNA double helix (Levinson et al. 1985). SSM previously has been implicated in a variety of short tandem duplication events. We have therefore examined the general features of SR-DNA to determine whether they are consistent with the expected consequences of SSM.

SSM Can Readily Explain Key Features of SR-DNA

The consequences of SSM can provide a coherent explanation for the origin and evolution of simple repetitive sequences in genomic DNA, including many of the

of purines or pyrimidines (B. F). A. Sequences from centromeric regions of yeast chromosomes III and XI (Bloom et al. 1982). Simple repetitive regions of the two sequences (underlined) have A + T contents of 93%. Deletions in various parts of these sequences result in loss of centromeric function. B. Noncoding, predominantly polypyrimidine sequence downstream from the polyadenylation site of a rat immunoglobulin kappa chain constant region gene (H. W. Sheppard and G. A. Gutman, unpublished data). Tandem repeats of (CT)n, (CTGTT)n, (CTCTT)n, and (CTCTTTT)n are emphasized. C. Mouse cDNA clone (Epplen et al. 1983) containing simple tandem repeats of CA, GA, GATA, GACA, and related motifs (underlined). Probable tandem duplications, each containing various simpler repeats, are indicated by arrows over the sequence. D. Sequences from mouse and Drosophila that cross-hybridize because they contain similar quasi-repeat units: 1 and 2, germ-line sequences mediating class-switching rearrangements in mouse immunoglobulins. Various perfect and imperfect 5-base repeats are emphasized; underlined regions are discussed below. Sequences are from the "S" region, 5' to the C region (Davis et al. 1980); 3: Quasi-repeat sequences isolated from Drosophila by cross-hybridization to mouse immunoglobulin class-switching sequences (Sakoyama et al. 1982). Spacing emphasizes 9-base quasi-repeat units of GAGCTGGG\(7/6\); the first 8 bases of this motif (underlined) closely resemble underlined motifs in the mouse clones shown above. E. Sequence from cytomegalovirus Colburn (Jeang and Hayward 1983) with alternating purine-pyrimidine quasi-repeat units of length 2. These quasi repeats precede a long perfect tandem reiteration of CA; purine-pyrimidine alternations are underlined. F. Long, predominantly homopyrimidine tract from intron of the Y gene, a member of the chicken ovalbumin family (Heilig et al. 1982). Tandem repeats of various lengths are emphasized (but only those involving \(\geq 8\) bases and with repeat units of length \(\geq 2\)). Perfect and near-perfect probable tandem duplications, each containing several repetitive motifs, are indicated by arrows; imperfections in these repeats are underlined.
repetitive tracts of satellite DNA commonly found in many eukaryotic genomes. The mechanistic basis for SSM was established >20 years ago (Fresco and Alberts 1960; Kornberg et al. 1964; Kornberg 1980, pp. 143-145). In its simplest form, SSM involves local denaturation and displacement of the strands of a DNA duplex followed by mispairing of complementary bases at the site of an existing short tandem repeat. The simplest consequences of this mispairing, when followed by replication or repair, can lead to insertions or deletions of one or several of the short repeat units. Figure 2A shows how mispairing during DNA replication could lead to an insertion or deletion. Figure 2B suggests a second possible mechanism in which mispairing of intact chromosomal DNA, followed by excision/repair, could lead to insertions or deletions (see also Flanagan et al. 1984).

On surveying a variety of published and unpublished SR-DNA sequences, we can discern several relevant and general features. We list some of these below and discuss their relationship to SSM.

First, tandem repeat units (repeated motifs) can vary in length from 1 to >10 bases, and any of the four nucleotides can participate. This may be related to the high probability of chance occurrence of short simple repeats. For example, in a completely random sequence, the probability of obtaining a 6-base run of a 1- or 2-base motif (such as AAAAAA or ACACAC) is 1/256, since there are 16 possible motifs and a probability of 1/4,096 for each run. Simple repeats that occur by chance may provide abundant raw material for expansion by SSM, as shown in figure 2. Once expanded, a short repeat should provide an even more efficient substrate for SSM, increasing the likelihood of additional slippage events. This is supported by observations that frequencies of spontaneous insertions and deletions in runs of [Al4 and [Al3 increase by more than an order of magnitude when the length of each of these runs is increased by a single base (Streisinger and Owen 1985). Also, we have observed that very long tandem repeats borne by coliphage M13 show extremely high frameshift frequencies, >1% in [CA]20 (G. Levinson and G. A. Gutman, unpublished data). Therefore, to the extent that SSM results in expansion of simple repeats, it might be expected to have a self-accelerating component.

Second, tandem repeat tracts containing motifs that differ by a single change (a substitution or size difference) are often found in close proximity and are often contiguous (Epplen et al. 1983; Levinson et al. 1985; H. W. Sheppard and G. A. Gutman, unpublished data; examples are shown in figs. 1B, 1C). Such a pattern can readily be understood as the consequence of multiple SSM events occurring before and after base-substitution events. A mutational change (substitution, insertion, or deletion) can create new repeat units from existing ones (e.g., a transition can change AAAAA to AAGAA), and subsequent SSM events that are likely to occur in an already repetitive region can then expand these new motifs as shown in figure 3; the result would be a new tandem repeat adjacent to the old one. Much of the variety of tandemly repeated motifs could be explained in this fashion.

Third, long repetitive tracts of certain motifs—including polypyrimidine tracts (polypurines on the complementary strand) and poly-CA tracts—are frequently observed. The prevalence of long polypyrimidine tracts (Straus and Birnboim 1976; Dodd and Straus 1982; Heilig et al. 1982; Sorge and Hughes 1982; Schmid and Shen 1985; as illustrated in figs. 1B, 1F) may be due to the combined influence of SSM and base substitutions (described above) plus the greater likelihood that base substitutions will be transitions rather than transversions (Fowler et al. 1974; Topal and
Fig. 2.—Generation of duplications or deletions by SSM between contiguous repeats. Small arrows indicate direction and starting point of DNA synthesis; colons indicate base pairing. A, 2-Base slippage in an AT-repeat during replication of a DNA duplex, followed by continued chain elongation. Slippage in the 3' → 5' direction (left panel) results in insertion of one AT unit; slippage in the other direction (right panel) results in deletion of one repeat unit. The deletion shown on the right results from excision of the unpaired repeat unit (asterisks) at the 3' end of the growing strand, presumably by the 3' → 5' exonuclease activity of DNA polymerase. B, The same slip occurring in intact duplex DNA. Mismatched regions form single-stranded loops, which may be targets for excision and repair. Results depend on where excision/repair events take place: excision of the shorter loop on the top strand, followed by repair synthesis using the lower strand as template, results in addition of one AT repeat unit as shown; other outcomes, including deletions, are also possible.

Fresco 1976; Brown et al. 1982; Holmquist 1983; Li et al. 1985, pp. 16–28, 43–54). Transitions, by definition, change pyrimidines to other pyrimidines and purines to other purines. Since transitions are more likely than transversions, it is (1) more probable that existing pyrimidine tracts will be maintained and (2) more likely that new simple pyrimidine repeats will arise by chance (and be subject to subsequent expansion) as base transitions accumulate. If SSM commonly generates longer repeats from shorter ones and base transversions are relatively infrequent, long polypyrmidine tracts, containing a variety of perfect and imperfect tandem repeats, would tend to result.

The prevalence of poly-CA tracts (Hamada et al. 1982a, 1982b; Jeang and Hayward 1983; Rogers 1983; Schmid and Shen 1985) can be explained in a similar fashion (Li et al. 1985, pp. 43–54). Methylated C residues are subject to deamination, causing a transition of C to T (Coulondre et al. 1978; Razin and Riggs 1980). Since ~90% of methylated C residues reportedly occur at 5' CG 3' nucleotides (Razin and Riggs 1980), this process would tend to increase the abundance of 5' TG 3' motifs, along with their complementary 5' CA 3' motifs. As we have argued for polypyrmidine tracts, it follows
FIG. 3.—Transformation of simple repeats into more complex ones by propagation of base substitutions. A base substitution (marked by circle) occurring in a simple repetitive region (poly-A/poly-T) can be propagated by SSM events. The result is the generation of a repetitive region containing a new motif (GA/CT), adjacent to the original poly-A/poly-T. In this example the base substitution is a transition, and the old and new repetitive regions together form a polypyrimidine/polypurine tract. Unpaired bases are indicated by asterisks.

that the increased frequency of TG/CA motifs would enhance the fortuitous occurrence of tandem repeats—and that these would be subject to subsequent expansion by SSM, generating long tracts of poly-CA.

The above explanation for the preferential expansion of polypyrimidine and poly-CA tracts should apply to other nonrandom patterns of base substitution, including patterns deriving from broadly relevant trends as well as those associated only with specific organisms. In the absence of adverse selection and under conditions that favor expansion, motifs that occur more frequently should be subject to a greater degree of expansion than those that occur infrequently, with a consequent increase in the length and abundance of corresponding tandem repeats.

Fourth, short simple repeats are often included within longer repeats (Appels and Peacock 1978; Brutlag 1980; Heilig et al. 1982; Miklos and Gill 1982; Singer 1982; Epplen et al. 1983; Miklos 1985; Walsh, accepted; see examples in figs. 1C, 1F). This can be understood in part as a consequence of mutational events creating new, longer motifs from tandemly arranged shorter ones, as discussed above. The juxtaposition
of closely related repeats can form a longer, more efficient substrate for duplication by SSM, resulting in the formation of a larger repeat unit. In addition, tracts of simple repeats, because of their ability to mispair, may be predisposed to long tandem duplications by unequal crossing-over (UCO) and other interhelical events, as discussed in more detail below.

SSM Has Been Invoked in Various Contexts

In early in vitro studies, Fresco and Alberts (1960) showed that the helix of double-stranded RNA can readily accommodate single-stranded loops of $\geq 1$ unpaired bases and showed by model building that DNA double helices should behave similarly. On the basis of these observations, they proposed that formation of short loops of unpaired bases by mispairing could lead to insertions or deletions (as well as substitutions). In other early studies, Kornberg and colleagues showed that double-stranded DNA oligomers, such as oligo-AT, can effectively prime the synthesis of high-molecular-weight reiterated DNA by *E. coli* DNA polymerase I in vitro (Kornberg et al. 1964; Kornberg 1980, pp. 143–145). Reactions were favored by elevated temperatures, with longer primers having higher temperature optima than shorter ones, observations that imply that disruption of normal base pairing is required for reiteration to occur. These workers proposed that repeated rounds of strand slippage combined with primer extension could explain these results.

Wells and colleagues (1967a, 1967b) extended these experiments to double-stranded oligomeric primers containing repeat units of 3 or 4 bases, showing that incubation of such oligomers at high temperature led to production of high-molecular-weight DNA. In every case, the tandem repeats in the polymers matched those in the oligomeric primers; e.g., a mixture of [TAGA]$_2$ plus [TATC]$_3$ primed the synthesis of poly-TAGA.

SSM in vivo has been invoked to explain small insertions and deletions of tandem repeat units in a variety of studies of spontaneous mutations in *E. coli*. In the oft-cited study by Streisinger et al. (1966), SSM was proposed as an explanation for spontaneous frameshift mutations in bacteriophage T4. More recently, a variety of studies have shown that short single-base runs (Pribnow et al. 1981; Levin et al. 1982; Owen et al. 1983; Streisinger and Owen 1985) or tandem repeats of other simple motifs (Farabaugh et al. 1978) are hot spots for frameshift mutations. Frameshift hot spots are not restricted to simple repeats, however; other hot spots may involve novel pairing configurations within each of the DNA strands of quasi-palindromic sequences (Ripley 1982; DeBoer and Ripley 1984). SSM has also been used to explain various features of eukaryotic DNA sequences, including tandem reiterations (Kornberg 1964; Kornberg et al. 1980, pp. 143–145; Jones and Kafatos 1982; Moore 1983; Rodakis et al. 1984; Tautz and Rcnz 1984a, 1984b), duplications and deletions (Efstradiatis et al. 1980) including coupled events (Flanagan et al. 1984), gene conversion (Slihtom et al. 1980), and illegitimate recombination and viral integration (Hasson et al. 1984).

Distinguishing between Intrahelical and Interhelical Events

Besides SSM events, UCO can also generate tandem duplications in DNA. UCO is, in fact, widely viewed as an important force in the generation and maintenance of multigene families as well as satellite DNA (Ohno 1970; Smith 1973, 1976; Anderson and Roth 1977, 1981; Kurnit 1979; Strickberger 1985, pp. 507–509, 757–760). These two mechanisms have important features in common. They both can generate duplications (and deletions) in DNA in a manner dependent on homologous base pairing
and, as a result, should both be self-accelerating for duplications. On the other hand, the mechanisms differ in that SSM is an intrahelical event, involving the two strands of a single DNA duplex, whereas UCO is an interhelical event, involving DNA molecules from two different chromosomes or sister chromatids. This places special constraints on UCO, since it can only take place during chromosome alignment in cell division and will be dependent on such factors as the rate of chiasma formation. SSM, on the other hand, ought to be free of such constraints and could potentially occur whenever unpaired loops form, during DNA repair as well as replication. SSM might therefore be expected to be an inherently more frequent event.

Walsh (accepted) has pointed out that another type of event involving crossovers within a chromatid can also occur; however, such events would always result in deletions—and so would tend to oppose the expansive potential of both SSM and UCO.

Another consequence of the intrahelical nature of SSM is the expectation that SSM should have an appreciable bias toward the duplication of shorter repeat units; if the initial event involves local melting and reannealing of the duplex, then a shorter slippage should be more likely than a longer one, since it distorts the normal configuration of the molecule less. Observed rates of SSM in vitro are consistent with this expectation; Wells et al. (1967b) found that elongation rates decreased considerably when the length of the repeat unit was increased from 2 to 4.

UCO, on the other hand, should be limited primarily by the total length of sequence available for unequal pairing—but with little regard for the degree of slippage required—since the misalignment takes place on a chromosomal rather than on a molecular scale. Computer modeling of UCO (Smith 1976), in fact, gave rise to a broad range of repeat-unit lengths with a mean of 18 but with no bias toward the shorter motifs.

If this analysis is correct, and if SSM is a ubiquitous process, one would expect to find, in otherwise unselected DNA sequences, evidence for the propagation in genomic DNA of the shortest repeat units, namely, runs of single-base motifs. We have performed computer-based analysis of mammalian DNA sequences and have obtained results that indicate that single-base repeats form longer runs in natural sequences (in intervening sequences specifically) than would be expected on a random basis. Figure 4 shows the frequency distribution of runs of a single base in natural sequences taken from 91 introns of 25 mammalian genes, compared with their pseudorandom counterparts (matched for base composition and length). It is clear that, for every size category above length 2, its representation in natural sequences is greater than that in the random sequences, a difference that is most striking in the greater-length categories. Thus, there is a substantial excess of longer runs of a single base in the natural sequences. These findings are an extension of those of Blaisdell (1983), who reported a “global non-randomness” in the 1-base runs of introns.

Therefore, in mammalian introns not chosen for their content of known repetitive sequences, some influence has driven single-base repeats to increase in length. We would propose, on the basis of arguments outlined above, that SSM, rather than UCO, is likely to be the mechanism involved—and is therefore likely to be a ubiquitous force influencing the evolution of DNA sequences.

SSM May Generate Large Duplications and Predispose DNA to Interhelical Events

For simplicity, the above discussion of the SSM mechanism has been limited to mispairing between tandem repeats (fig. 2). However, as a genomic region becomes
increasingly simple and repetitive, the probability that noncontiguous sequences will mispair should also increase. Such noncontiguous (although still intrahelical) mispairing events could lead to larger duplications, deletions, palindromes, and other rearrangements. A hypothetical duplication event involving noncontiguous SSM is shown in figure 5.

Another likely consequence of noncontiguous SSM events is the deletion of sequences between direct repeats, a common occurrence both in vitro (Kunkel 1985) and in vivo (Livneh 1983; Owen et al. 1983). Such deletion events would have the effect of joining two nonadjacent repetitive tracts into a single continuous one. This same process has also been invoked to explain putative coupled deletion/duplication events that appear to have occurred within human alpha-immunoglobulin genes (Flanagan et al. 1984).

Analysis of eukaryotic sequences has led to the suggestion that regions of SR-DNA may be hot spots for interhelical events, such as gene conversion (Slighom et al. 1980) and illegitimate recombination (Hasson et al. 1984). If this is so, then expansion of short repetitive sequences by SSM would be expected to increase the likelihood of these types of events. Such an effect could be explained by at least two factors. First, longer repetitive regions would provide a much more efficient substrate for the complementary but unequal pairing required of UCO, as we have already mentioned; in fact, simple tandem repeats have been implicated as hot spots for UCO events (see, e.g., Jeffreys et al. 1985). Second, the single-stranded regions arising during SSM could
directly encourage interhelical events. Single-stranded loops would presumably stimulate branch migration, a process that has been implicated in homologous recombination (Lee et al. 1970; Warner et al. 1979). Single-stranded loops should also be targets for excision/repair (Hanawalt et al. 1979; Kornberg 1980, pp. 340-343; Glickman 1981; Grossman 1981; Kramer et al. 1982, 1984; Lu et al. 1983; Flanagan et al. 1984), and this process could generate the free ends of DNA that might also participate in either legitimate or illegitimate recombination events (i.e., events involving extensive or limited sequence identity, respectively [Radding 1978]).

Studies with S1 nuclease have provided direct evidence that DNA sequences containing short tandem repeats are prone to the spontaneous formation of transitory single-stranded regions (Hentschel 1982; Mace et al. 1983; Weintraub 1983; Hamada et al. 1984a). It has been suggested that relief of the torsion of supercoiling may drive the formation of such regions (Nickol and Felsenfeld 1983). If such single-stranded
regions are generally characteristic of SR-DNA, they could predispose DNA to both SSM events (potentially contributing to the self-accelerating character of this process) as well as to the interhelical events discussed above.

Features of Satellite DNA

Satellite DNA sequences make up large proportions (e.g., as much as 44% of the nuclear genome of some higher plants [Ingle et al. 1973]) of eukaryotic genomes. Despite its abundance, the origin and functional significance (if any) of satellite DNA, though much discussed, is not presently understood. All of the general features of SR-DNA discussed above are common to many satellites, suggesting that SSM might play a major role in the evolution of these structures. Relevant features of satellite sequences have been reviewed by Appels and Peacock (1978), Brutlag (1980), Miklos and Gill (1982), Singer (1982), Miklos (1985), and Walsh (accepted).

Many satellite DNAs contain high proportions of very simple repetitive motifs; for instance, crab satellite contains poly-AT sequences (Hamori 1975), and the snake satellite that we previously have studied (Levinson et al. 1985) contains interspersed GATA and GACA motifs. Satellite DNA can also contain tracts of closely related repeat motifs, either interspersed or in tandem arrays, as in the case of the snake satellite cited above. In addition, an apparently haphazard collection of single-base runs within two repeat units of the 1.688-g/cc satellite of *Drosophila* has been described (Carlson and Brutlag 1979; Hsieh and Brutlag 1979; Miklos and Gill 1981). There are also numerous examples of satellite sequences with repeat units that themselves contain shorter simple repeats (Appels and Peacock 1978; Brutlag 1980; Miklos and Gill 1982; Singer 1982; Epplen et al. 1983; Miklos 1985; Walsh, accepted).

In addition to simple tandem repeats, satellite sequences can also include self-complementary quasi-palindromic motifs, which may promote frameshifts (Ripley 1982; DeBoer and Ripley 1984) as well as DNA repair and hence might also play a role in the expansion of satellite DNA. Examples of self-complementary satellite motifs, including blocks of alternating pyrimidines and purines (Rosenberg et al. 1978), can be found in the data of Singer (1982).

The presence within some satellite sequences of repeat units much longer than those that we have been discussing is more difficult to reconcile with SSM events, however. Mouse satellite, for instance, contains a predominant repeat unit some 240 bp in length, together with others 120 bp and 480 bp long (Southern 1975), and the presence of the latter repeats has been interpreted as evidence for the participation of UCO events in the generation of this satellite DNA. However, Southern calculated that, on the basis of his estimated rates of recombination, UCO events would be >10 times too slow to account for the generation of the 240-bp units and suggested that other mechanisms (including SSM) may have been responsible.

Thus, our view that simple repeats may be prone to expansion by both short tandem duplications via SSM and longer tandem duplications by UCO and other interhelical events may be illustrated well by satellite sequences. However, although SSM may play an important role, the precise mechanisms by which satellite sequences are expanded to high copy numbers remain unclear.

What Forces Could Account for the Accumulation of SR-DNA?

The proposed mechanisms for SSM events can generate *either* insertions or deletions, depending on the manner in which the mispaired structure is resolved (Fresco
and Alberts 1960). Some data on the relative frequencies of insertions versus deletions is available from studies of spontaneous mutations in bacterial genes. In some cases, insertion rates have been shown to be higher than those of deletions; of 94 spontaneous frameshifts within tandem repeats of CTGG, 76 were probable insertions and 18 were probable deletions, representing an excess of insertions of ~4:1 (Farabaugh et al. 1978). On the other hand, spontaneous frameshifts within various runs of a single base were found to be skewed toward deletions rather than insertions; deletion:insertion ratios ranging from 2:1 to 4:1 were observed in bacteriophage T4, and ratios of ~5:1 were predicted on thermodynamic grounds (Streisinger and Owen 1985). Bacterial frameshifts in long runs of 2-base motifs may also be skewed toward deletions: in a 40-bp poly-CA tract borne by bacteriophage M13, we have observed spontaneous deletion:insertion ratios (of single 2-base repeat units) of ~3:1 (G. Levinson and G. A. Gutman, unpublished results).

If deletions are occurring more frequently than insertions, it is difficult to explain how SR-DNA could progressively accumulate in the many eukaryotic contexts where it has been seen, particularly in satellite DNA. Two general possibilities can be invoked. First, multicellular eukaryotes may have higher intrinsic proportions of insertions than bacteria. Bacteria are subject to high selective pressure for rapid replication and cell division, and so the genetic apparatus might have evolved a bias toward deletions vis-à-vis insertions, in order to minimize genome size and maximize replication rate. Genome size may be less critical in multicellular eukaryotes, and their genetic apparatus may tend to favor insertions over deletions, either by generating insertions more frequently than deletions or by repairing insertion heteroduplexes less efficiently. Second, selective pressures may exist that encourage the long-term retention of SR-DNA. In either case, if there does exist a bias toward production or retention of insertions, then selection against the duplicated sequences would be required to prevent SR-DNA from continuously accumulating; such selection would certainly be expected to be the dominating factor within coding sequences. However, regions not subject to such negative selection would be expected to rapidly expand their repetitive sequences, potentially giving rise to large quantities of what has been termed “junk DNA” (Ohno 1972) or “selfish DNA” (Doolittle and Sapienza 1980; Orgel and Crick 1980).

If noncontiguous SSM events preferentially delete nonrepetitive sequences between two direct repeats (Livneh 1983; Owen et al. 1983; Flanagan et al. 1984; Kunkel 1985), as has been suggested, this could also, in effect, create a local bias toward expansion of repetitive elements; repetitive elements could be duplicated or deleted by SSM, but nearby nonrepetitive sequences would be preferentially deleted. Thus, in a region of DNA under no selective constraint except to maintain its overall length, nonrepetitive sequences would be systematically replaced by repetitive ones. Simple repeats might therefore constitute a natural ground state of unselected DNA, analogous to what has been suggested by Smith (1976) on the basis of his analysis of UCO. Some other influence, either selective or stochastic, would still be required to explain the wholesale expansion of SR-DNA evident in satellite sequences.

What selective forces could act to conserve simple repetitive sequences? One possibility arises from the finding that centromeric function may be dependent on the presence of simple repeats (Bloom et al. 1982). Another is suggested by the association between SR-DNA and a variety of genetic regulatory elements. Examples include the imperfect joining and rearrangement of simple repetitive gene segments in both immunoglobulin and T-cell receptor gene families; these are important elements in the generation of antibody and T-cell receptor diversity (Kronenberg et al. 1986). Another
example is the repetitive element (shown in fig. 1) involved in class switching of mammalian immunoglobulin heavy-chain genes (see Ohno 1981). Repetitive sequences have also been found as part of the enhancer associated with one of the mouse major-histocompatibility-complex genes (Gillies et al. 1984). These authors found two polypyrimidine tracts totaling 95 bp, a polypurine tract of 71 bp, and an alternating purine/pyrimidine tract (a structure associated with the ability to form Z-DNA) of 164 bp, all closely associated with the core enhancer elements and all present on the most active fragment that they isolated. Although these authors were not able to show that any of these repetitive elements was biologically active when isolated from the other elements, Hamada et al. (1984b) identified a simple repetitive sequence (poly-CA) that could function by itself as an effective transcriptional enhancer in transitory in vivo assays.

Thus, simple repetitive sequences, at least those near expressed genes, might provide raw material for the evolution of regulatory elements. The ability to function in this manner may arise from the special structural properties of SR-DNA, some of which are a consequence of the highly skewed base composition of such sequences. Poly-CG, for instance, failed to act as an enhancer in the system designed by Hamada et al. (1984b) whereas poly-CA was effective; the former is a much more stably hydrogen-bonded duplex than the latter—or than any other sequence not consisting totally of G/C pairs. In the case of an alternating purine/pyrimidine sequence, its ability to form Z-DNA might also confer on it the capability to function in some regulatory capacity, but the failure of poly-CG to do so in Hamada's system implies that other physical properties of the sequences may also play a decisive role. One intriguing possibility is that the single-stranded loops associated with mispairing in certain simple repeats (Hentschel 1982; Mace et al. 1983; Weintraub 1983; Hamada et al. 1984a) might function in a regulatory capacity as a result of their unique physical properties.

Evolution of SR-DNA: An Overview

We have argued that SSM events can account for many of the features characteristic of simple repetitive DNA and are therefore likely to have played a major role in the origin and evolution of the latter. We can summarize our views on the development of SR-DNA as follows:

1. Short, simple tandem repeats that arise by chance in DNA sequences can be expanded by SSM events into longer tandem repeats.
2. Mutational changes (base substitutions, insertions, or deletions) can create new motifs that may be propagated by additional SSM events; this would give rise to tandem or interspersed repeats of closely related motifs. Also, nonrandom patterns of base substitution would increase the length and abundance of particular simple repeats, including polypyrimidine and poly-CA tracts.
3. As repetitive regions become longer, the probability of noncontiguous SSM increases, increasing the possibility of longer tandem duplications. Such events may also tend to delete nonrepetitive sequences between repeats that are capable of mispairing, thereby increasing the length and homogeneity of repetitive tracts.
4. As regions of SR-DNA expand, they may be predisposed to more rapid expansion by means of UCO or other interhelical events by virtue of their mispairing potential and single-stranded character. This may generate longer tandem duplications that would contain within them the shorter tandem repeats originally expanded by SSM.
5. The net evolutionary result of such events will be critically dependent on the relative rates of SSM, point mutations, UCO, and other processes that can alter DNA structure. The overall expansion of SR-DNA will also be influenced by the degree to which SSM and UCO events are intrinsically biased toward insertions or deletions and by the (unknown) selective forces that may act to retain or eliminate repetitive regions.

Acknowledgments

The authors thank H. Tucker for statistical analysis of single-base runs; W. M. Fitch, M. B. Frank, A. Konopka, J. E. Manning, R. Sandri-Goldin, S. J. Sharp, R. C. Warner, and A. C. Wilson for helpful criticisms of the manuscript; and K. P. Bertrand, M. Guiltinan, and M. Zarahovic-Radic for stimulating discussions and suggestions in the course of its preparation. This work was supported by U.S. Public Health Service grants AI-14774 and AI-21366 and an award from the Chancellor's Patent Fund (University of California, Irvine). G.L. was supported by National Institutes of Health Research Service Award HD07029 and Earle C. Anthony and Monsanto Company Fellowships. This work was initiated while G.L. was affiliated with the Department of Development and Cell Biology and the Developmental Biology Center of the University of California, Irvine.

APPENDIX A

Computer Analysis of Single-Base Repeats in Mammalian Genes

Genes to be analyzed were selected from an alphabetical listing of mammalian DNA genomic sequences in the GenBank database (obtained from Bolt Beranek and Newman, Inc., Cambridge, Mass.), release 24.0, dated September, 1984. Sequences were chosen from diverse gene families to avoid biasing the results toward highly represented genes (e.g., globins and immunoglobulins). A computer program called SIMPLDNA (G. Levinson, M. K. Nistanaki, L. Howell, S. Anderson, and G. A. Gutman, unpublished data), written in TURBO-PASCAL for the IBM-PC, was used to determine the frequency distribution of the lengths of runs of a single base. Analysis was performed on all introns and exons of each analyzed gene, except that untranslated exons and partial intron or exon segments <20 bases in length were excluded from the analysis. Introns and exons were identified by comment lines in the GenBank files. The GenBank file, organism, and protein names of the analyzed gene sequences are as follows: BOVGH, bovine growth hormone; BOVPS1-5, bovine opsin; BOVOMC3-5, bovine proopiomelanocortin; DOGINS, dog insulin; GOTHBAI, goat adult alpha-i-globin; HAMVIM1-7, hamster vimentin; HUM1AT1-4, human alpha-1 antitrypsin; HUMACTCA1-4, human alpha-cardiac actin; HUMAPOAI 1, human apolipoprotein A-I; HUMCMYCB2-3 human c-myc oncogene; HUMGLYCA1-4, human glycoprotein, alpha-subunit; HUMIFNG, human immune interferon; HUMMETII, human metallothionein II; HUMMH, human class 1 transplantation antigen (HLA); HUMMHDRS1-2, human HLA-DR alpha-chain (chain p34); HUMPLA, human placental lactogen hormone; HUMPTH1-2, human parathyroid (PTH); HUMTBBM40, human beta-tubulin; MUSAMY1A3-4, mouse alpha-amylose-1; MUSFOL1-6, mouse dihydrofolate reductase; MUSIDC10, mouse immunoglobulin germ-line d-j-c region: mu; RABHBB1A1, rabbit beta-1 globin; RATAVP1-2, rat arginine vasopressin-neurophysin precursor; RATCAG11-12, rat gamma casein; and RATCYC, rat (Sprague-Dawley) cytochrome C.

For comparison, an analogous, pseudorandom sequence was computer generated for each of the 91 introns surveyed, each of which had the same length and base composition as its natural counterpart. These 182 natural and pseudorandom sequences
were then analyzed for single-base runs with SIMPLDNA. The results are shown in figure 4 and are discussed in the text.

The percentage of all nucleotides in a given sequence that form monomer runs of length $\geq 2$ were also compared, to determine whether a skewed base composition would generate the same degree of total repetition in the pseudorandom sequences. In fact, the percentages of nucleotides in such runs for natural and computer-generated sequences were very similar, the ratio of introns versus their random counterparts being 1.04; it was the size distribution of these runs that was different, as discussed in the text.

This approach is a conservative one. Since, in our analysis, we are comparing natural sequences with their pseudorandom counterparts, repetitiveness resulting solely from a skewed base composition is compensated for; for example, if a particular sequence consists entirely of a run of a single base, it would appear to our analysis as being no more repetitive than its pseudorandom counterpart, even though it is a perfect single-base repeat. Thus, we may be underestimating the degree to which mammalian introns are biased toward containing simple repeats.

LITERATURE CITED


CATO, A. C. B., S. GEISSE, M. WENZ, H. M. WESTPHAL, and M. BEATO. 1984. The nucleotide sequences recognized by the glucocorticoid receptor in the rabbit uteroglobin gene region are located far upstream from the initiation of transcription. EMBO J. 3:2771–2778.


JEANG, K.-T., and G. S. HAYWARD. 1983. A cytomegalovirus DNA sequence containing tracts


walking shows a highly homologous repetitive sequence present in all the centromere regions of fission yeast. EMBO J. 5:1011-1021.


WALSH, B. Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. Genetics (accepted).


WALTER M. FITCH, reviewing editor

Received April 23, 1986; revision received November 5, 1986.
The Rate of Synonymous Substitution in Enterobacterial Genes Is Inversely Related to Codon Usage Bias

Paul M. Sharp and Wen-Hsiung Li
Center for Demographic and Population Genetics, University of Texas

Gene sequences from *Escherichia coli*, *Salmonella typhimurium*, and other members of the Enterobacteriaceae show a negative correlation between the degree of synonymous-codon usage bias and the rate of nucleotide substitution at synonymous sites. In particular, very highly expressed genes have very biased codon usage and accumulate synonymous substitutions very slowly. In contrast, there is little correlation between the degree of codon bias and the rate of protein evolution. It is concluded that both the rate of synonymous substitution and the degree of codon usage bias largely reflect the intensity of selection at the translational level. Because of the high variability among genes in rates of synonymous substitution, separate molecular clocks of synonymous substitution might be required for different genes.

Introduction

Comparative studies of protein sequence data have provided ample evidence that the rate of molecular evolution is inversely related to the degree of selective constraint on the sequences or sites involved (Dayhoff 1972; Kimura 1983). Thus, given that there is much variation among proteins in the functional stringency of the precise amino acid sequence, it was not surprising to discover that the rate of nonsynonymous (i.e., amino acid-replacing) nucleotide substitution varies greatly among genes (Miyata et al. 1980; Kimura 1983). On the other hand, it was commonly thought that most synonymous mutations are subject to no strong selective constraints and that therefore the rate of synonymous substitution should not vary much among genes. Analyzing the then available DNA sequence data, Miyata et al. (1980) found that this was indeed the case, and they suggested that synonymous substitutions from different genes could be pooled to serve as a molecular clock. However, subsequent analysis of a larger number of mammalian genes has revealed that the synonymous rate varies considerably among genes (Li et al. 1985b). Although there was a positive correlation between the synonymous and nonsynonymous rates in the same gene, this correlation could explain only part of the variation in the synonymous rate (Graur 1985; Li et al. 1985b).

On the other hand, a large body of evidence has been gathered that demonstrates that the use of alternative synonymous codons is far from random (Grantham et al. 1981; Ikemura 1985). The pattern of codon usage is species specific (Grantham et al. 1981), but within each species there is considerable variation in the degree of codon usage bias. For example, in *Escherichia coli* and the yeast *Saccharomyces cerevisiae* the bias is much stronger in highly expressed genes than in moderately and lowly

1. Key words: synonymous codon usage, synonymous substitution rate, molecular clock, translational efficiency, *Escherichia coli*, *Salmonella typhimurium*.

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

© 1987 by The University of Chicago. All rights reserved.
0737-4038/87/0403-0002$02.00

222
expressed genes (Bennetzen and Hall 1982; Gouy and Gautier 1982; Sharp and Li 1986; Sharp et al. 1986). Selective differences among synonymous codons have been hypothesized, largely related to the relative abundance of isoaccepting tRNAs and their interaction with different codons (Ikemura 1981, 1985; Grosjean and Fiers 1982). In *E. coli* and yeast, very highly expressed genes use almost exclusively the small group of codons that are presumed to be optimal for efficient translation (Ikemura 1981, 1985; Bennetzen and Hall 1982). Apparently, the large variation in codon usage bias in *E. coli* and yeast genes reflects variation in selective constraints mediated through the process of translation.

Possibly the variation among genes in the rate of synonymous substitution and in the degree of codon usage bias are two aspects of the same phenomenon, i.e., both may reflect the variation among genes in selective constraints on synonymous-codon usage. If this is true, then the rate of synonymous substitution should be inversely related to the degree of codon usage bias. A preliminary study based on a few genes (Ikemura 1985) has suggested that this is true. Here we use recently determined DNA sequence data from *E. coli*, *S. typhimurium*, and a few related species to demonstrate that a significant negative correlation indeed exists between the rate of synonymous nucleotide substitution and the degree of codon usage bias.

**Material and Methods**

DNA sequence data are available for 23 pairs of homologous genes from *Escherichia coli* and *Salmonella typhimurium* (table 1) and for several pairs from some related species (table 2). Estimated numbers of synonymous and nonsynonymous nucleotide substitutions between genes were calculated by a method that takes account of both the degree of degeneracy of nucleotide sites and the different rates of transitions and transversions (Li et al. 1985b).

The degree of synonymous-codon usage bias was measured by the "codon adaptability index" (CAI; Sharp and Li 1987), which estimates the extent of bias toward codons that are known to be favored in highly expressed genes. Each codon is assigned a relative "adaptiveness" value according to its frequency of use in a species-specific reference set of very highly expressed genes, and the CAI for a gene is then calculated as the geometric-mean relative adaptiveness of its codons. A CAI value of 1.0 indicates that the gene in question contains only optimal codons and thus has the greatest possible degree of bias. A value close to zero indicates extensive use of codons that normally are rare. Examination of very weakly expressed *E. coli* genes suggests (1) that the frequencies of these rare codons are usually even lower than would be expected on the basis of random codon usage (Sharp and Li 1986) and (2) that the observed CAI values for *E. coli* genes are all higher than the CAI value (~0.17) for a sequence in which all 61 sense codons are equiprobable. *Escherichia coli* and *S. typhimurium* exhibit similar patterns of codon bias (Ikemura 1985), presumably because they are so closely related that similar constraints operate on codon usage in the two species. A reference set of 27 very highly expressed *E. coli* genes (mainly ribosomal protein, elongation factor, and outer-membrane-protein genes; see Sharp and Li 1986) was used to calculate CAI values for both the *E. coli* and *S. typhimurium* genes (table 1). For each gene, the CAI values for the two species are quite similar.

**Results**

Since the divergence time is the same for all orthologous pairs of *Escherichia coli* and *Salmonella typhimurium* genes, we can compare the relative rates of evolution
Table 1
Comparison of Genes: *Escherichia coli* and *Salmonella typhimurium*

<table>
<thead>
<tr>
<th>Gene</th>
<th>No. of Codons</th>
<th>Mean $K_s$</th>
<th>Mean ± SE $K_s$</th>
<th>CAI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$K_s$</td>
<td></td>
<td><em>E. coli</em></td>
</tr>
<tr>
<td>trpA</td>
<td>267</td>
<td>0.08</td>
<td>1.77 ± 0.33</td>
<td>0.34</td>
</tr>
<tr>
<td>cheY</td>
<td>128</td>
<td>0.01</td>
<td>1.49 ± 0.31</td>
<td>0.30</td>
</tr>
<tr>
<td>trpC</td>
<td>451</td>
<td>0.07</td>
<td>1.39 ± 0.18</td>
<td>0.31</td>
</tr>
<tr>
<td>tar</td>
<td>552</td>
<td>0.12</td>
<td>1.37 ± 0.14</td>
<td>0.32</td>
</tr>
<tr>
<td>aroC</td>
<td>280</td>
<td>0.04</td>
<td>1.27 ± 0.17</td>
<td>0.25</td>
</tr>
<tr>
<td>aroA</td>
<td>425</td>
<td>0.07</td>
<td>1.26 ± 0.14</td>
<td>0.36</td>
</tr>
<tr>
<td>pabA</td>
<td>186</td>
<td>0.08</td>
<td>1.24 ± 0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>dnaG</td>
<td>580</td>
<td>0.08</td>
<td>1.18 ± 0.11</td>
<td>0.27</td>
</tr>
<tr>
<td>hisIE</td>
<td>202</td>
<td>0.06</td>
<td>1.11 ± 0.17</td>
<td>0.39</td>
</tr>
<tr>
<td>trpE</td>
<td>519</td>
<td>0.07</td>
<td>1.06 ± 0.10</td>
<td>0.36</td>
</tr>
<tr>
<td>trpD</td>
<td>530</td>
<td>0.02</td>
<td>1.06 ± 0.10</td>
<td>0.34</td>
</tr>
<tr>
<td>trpB</td>
<td>396</td>
<td>0.02</td>
<td>1.03 ± 0.11</td>
<td>0.41</td>
</tr>
<tr>
<td>crp</td>
<td>209</td>
<td>0.00</td>
<td>0.89 ± 0.16</td>
<td>0.49</td>
</tr>
<tr>
<td>orf1$^{ab}$</td>
<td>108</td>
<td>0.04</td>
<td>0.84 ± 0.17</td>
<td>0.35</td>
</tr>
<tr>
<td>ilvY$^a$</td>
<td>257</td>
<td>0.01</td>
<td>0.72 ± 0.10</td>
<td>0.34</td>
</tr>
<tr>
<td>metB</td>
<td>385</td>
<td>0.02</td>
<td>0.57 ± 0.06</td>
<td>0.34</td>
</tr>
<tr>
<td>rpoD</td>
<td>612</td>
<td>0.01</td>
<td>0.49 ± 0.05</td>
<td>0.58</td>
</tr>
<tr>
<td>ilvM</td>
<td>86</td>
<td>0.03</td>
<td>0.47 ± 0.12</td>
<td>0.23</td>
</tr>
<tr>
<td>gltA$^a$</td>
<td>71</td>
<td>0.07</td>
<td>0.39 ± 0.12</td>
<td>0.64</td>
</tr>
<tr>
<td>ompA</td>
<td>345</td>
<td>0.04</td>
<td>0.35 ± 0.05</td>
<td>0.77</td>
</tr>
<tr>
<td>metJ</td>
<td>104</td>
<td>0.01</td>
<td>0.29 ± 0.08</td>
<td>0.41</td>
</tr>
<tr>
<td>rpoB$^a$</td>
<td>958</td>
<td>0.01</td>
<td>0.31 ± 0.03</td>
<td>0.63</td>
</tr>
<tr>
<td>rpsU</td>
<td>70</td>
<td>0.00</td>
<td>0.04 ± 0.03</td>
<td>0.73</td>
</tr>
<tr>
<td>5' tar</td>
<td>280</td>
<td>0.20</td>
<td>1.36 ± 0.21</td>
<td>0.31</td>
</tr>
<tr>
<td>3' tar</td>
<td>235</td>
<td>0.01</td>
<td>1.11 ± 0.17</td>
<td>0.34</td>
</tr>
</tbody>
</table>


* Partial sequence.
* Open reading frame upstream of pyrE.

in different genes (table 1). Interestingly, there is a large range in synonymous-substitution rates among genes. The two very highly expressed genes—rpsU and ompA—have both the greatest bias in codon usage and a very low degree of synonymous divergence (table 1). Three other genes with a high codon bias—rpoB, rpoD, and gltA—also show comparatively low rates of synonymous substitution. Among genes with a low codon bias (i.e., those in which CAI < 0.4), there is considerable variability in the synonymous-substitution rate, $K_s$; but high $K_s$ values predominate among the longer sequences (i.e., those in which I > 200), which have a smaller SE for the estimates of $K_s$.

$K_s$ has been plotted against the degree of codon bias (fig. 1). There is a significant negative linear correlation (coefficient 0.68, $P < 0.01$) between these two statistics, confirming that genes with a more extreme synonymous-codon bias undergo synonymous substitution at a slower rate. Among the genes with a low codon bias (CAI
Table 2

Comparison of Genes: Species of Enterobacteriaceae

<table>
<thead>
<tr>
<th>COMPARISON</th>
<th>CAI^a</th>
<th>K_s^b</th>
<th>K_Å^b</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. SPECIES PAIRS WITH A DIVERGENCE TIME EQUAL TO THAT BETWEEN Escherichia coli AND Enterobacter aerogenes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>E. aerogenes</em> ompA gene vs.:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>E. coli</em>...............</td>
<td>0.77</td>
<td>0.45</td>
<td>0.09</td>
</tr>
<tr>
<td><em>Shigella dysenteriae</em>........</td>
<td>0.75</td>
<td>0.51</td>
<td>0.10</td>
</tr>
<tr>
<td><em>Salmonella typhimurium</em>........</td>
<td>0.71</td>
<td>0.53</td>
<td>0.09</td>
</tr>
<tr>
<td><em>Klebsiella pneumoniae</em> pabA gene vs.:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>E. coli</em>...............</td>
<td>0.28</td>
<td>1.77</td>
<td>0.13</td>
</tr>
<tr>
<td><em>S. typhimurium</em>........</td>
<td>0.29</td>
<td>1.26</td>
<td>0.12</td>
</tr>
<tr>
<td><em>K. pneumoniae</em> trpA gene vs.:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>E. coli</em>...............</td>
<td>0.34</td>
<td>1.61</td>
<td>0.08</td>
</tr>
<tr>
<td><em>S. typhimurium</em>........</td>
<td>0.32</td>
<td>1.80</td>
<td>0.11</td>
</tr>
<tr>
<td><strong>B. SPECIES PAIRS WITH A DIVERGENCE TIME EQUAL TO THAT BETWEEN <em>E. coli</em> AND Serratia marcescens</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>S. marcescens</em> lpp gene vs.:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>E. coli</em>...............</td>
<td>0.85</td>
<td>0.33</td>
<td>0.07</td>
</tr>
<tr>
<td><em>S. marcescens</em> pabA gene vs.:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>E. coli</em>...............</td>
<td>0.28</td>
<td>1.47</td>
<td>0.16</td>
</tr>
<tr>
<td><em>S. typhimurium</em>........</td>
<td>0.29</td>
<td>1.57</td>
<td>0.19</td>
</tr>
<tr>
<td><em>K. pneumoniae</em>........</td>
<td><em>...</em></td>
<td>1.88</td>
<td>0.17</td>
</tr>
<tr>
<td><em>S. marcescens</em> trpG gene vs.:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>E. coli</em>...............</td>
<td>0.33</td>
<td>2.56</td>
<td>0.14</td>
</tr>
<tr>
<td><em>S. dysenteriae</em>........</td>
<td>0.37</td>
<td>1.74</td>
<td>0.15</td>
</tr>
<tr>
<td><em>S. typhimurium</em>........</td>
<td>0.31</td>
<td>1.60</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Footnote.—References for sequences are as in table 1.

^a Computed using a reference set of very highly expressed genes from *E. coli*.

^b K_s and K_Å = number of synonymous and nonsynonymous substitutions, respectively, between species.

< 0.4), the two with the lowest K_s values are rather short, and so the low K_s values could have arisen from sampling errors. The *ilvM* gene, which has a K_s value much lower than would be expected on the basis of its CAI value, is a short, previously unrecognized gene sandwiched between *ilvG* and *ilvE* and contains within its coding sequence a promoter for the *ilvE* gene (Lopes and Lawther 1986). The need to conserve this regulatory sequence may serve as an extra constraint reducing K_s in *ilvM*.

Sequence data are available for some genes from species of Enterobacteriaceae other than *E. coli* and *S. typhimurium*, notably *Klebsiella pneumoniae*, *Enterobacter aerogenes*, and *Serratia marcescens*. *Klebsiella pneumoniae* and *E. aerogenes* are considered to be more closely related to each other than either is to *E. coli* (Brenner 1984), and so the divergence times between *E. coli* (or *S. typhimurium*) genes and *K. pneumoniae* or *E. aerogenes* genes should be equivalent. *Serratia marcescens* is considered to be more distantly related to all of the above species (Brenner 1984), and so any comparisons between *S. marcescens* and the other species should involve equivalent time scales. Estimates of nucleotide substitution among these species are detailed in table 2. For the genes with low CAI values (i.e., *pabA*, *trpA*, and *trpG*), the K_s values are generally high—and near the level of saturation of synonymous substitutions. Again the genes with a high synonymous-codon bias (*ompA* and *lpp*) show a much
lower synonymous divergence. These data are entirely consistent with those in table 1 and figure 1, bearing in mind the longer divergence times.

Discussion

The correlation coefficient computed in figure 1 should be taken with caution because the estimated $K_S$ values may not be reliable when the true values are considerably higher than one (Li et al. 1985b). However, a significant correlation between CAI and $K_S$ is observed even if one uses the observed proportion of differences at synonymous sites instead of the $K_S$ value. For simplicity, we have assumed a linear correlation (fig. 1). We tried various transformations (e.g., logs, squares) of the axes, but the correlation coefficients obtained were very similar to one another.

Comparison of the rates of nucleotide substitution at noncoding sites (e.g., pseudogenes), coding but degenerate sites (i.e., sites at which synonymous mutations can occur), and amino acid–determining sites has revealed an inverse relationship between the rate and the degree of functional constraint (Kimura 1983; Li et al. 1985a). Here we extend that observation to synonymous sites in different genes that appear to be under different degrees of constraint, as interpreted on the basis of the observed degree of synonymous-codon bias. This confirms the preliminary observations of Ikemura (1985), which were made on the basis of data for only a few genes.

Horizontal transfer of genes between *Escherichia coli* and *Salmonella typhimurium* would produce gene pairs with surprisingly low synonymous divergence for a particular degree of codon bias. In figure 1 a few genes seem to be in this category,
but the two outstanding examples can be explained by the small number of codons examined (metJ) and/or extra sequence constraints (ibvM), so that interspecific exchange does not appear to be an important confounding factor here.

Among mammalian genes there is a tendency for genes with a high nonsynonymous rate, $K_A$, to have a high $K_S$ (Graur 1985; Li et al. 1985). Similarly here the correlation of $K_A$ and $K_S$ for the E. coli–S. typhimurium comparison is 0.57. Whereas $K_A$ is influenced by selection at the protein level, $K_S$ is presumed to be influenced by selection on synonymous-codon usage, i.e., at the level of translation. The correlation of $K_A$ and $K_S$ indicates that, among the genes studied, those that are highly expressed and tend to have a high CAI and a low $K_S$ also tend to encode conserved proteins. However, a direct relationship between protein sequence constraint and codon bias seems unlikely, since the (negative) correlation of $K_A$ and CAI is only 0.36. The importance of the precise amino acid sequence to protein function—and hence the degree of sequence conservation (and $K_A$)—can vary along a peptide. In contrast, the synonymous-codon composition is a property of an mRNA as a whole, so that the degree of codon bias (and hence $K_S$) should be comparatively uniform along a gene. This is dramatically illustrated by the tar gene. The amino-terminal half of the tar-gene product has diverged more rapidly than any other protein examined here, whereas the majority of the carboxy-terminal half is quite conserved (table 1, bottom). In contrast to the rate of amino acid replacement (reflected in $K_A$), $K_S$ is very similar in the two halves of the gene (table 1).

It has been suggested that in some cases the level of gene expression is modulated evolutionarily by the selection of rare codons to reduce the rate of translation (see, e.g., Grosjean and Fiers 1982; Konigsberg and Godson 1983). On the basis of this assumption it would be expected that this constraint on codon usage would reduce $K_S$, just as selection for optimal codons in highly expressed genes reduces the rate. In particular, it has been reported that the dnaG, lacI, trpR, and araC genes of E. coli have an excess of rare codons, a situation that has been explained as a mechanism to maintain low expression (Konigsberg and Godson 1983). We have shown elsewhere (Sharp and Li 1986) that these genes do not have significantly more rare codons than do a large number of other E. coli genes expressed at moderate to low levels. From the data presented here (table 1, fig. 1) it can be seen that dnaG and araC are accumulating synonymous substitutions at a rate typical of genes with a low codon bias. This suggests that the incidence of rare codons, yielding a low CAI, results from a comparative lack of negative (purifying) selection rather than from the presence of positive selection.

The approximate time ($t_{ES}$) of divergence of the E. coli and S. typhimurium lineages has been estimated previously by using rRNA molecular clocks (Hori and Osawa 1978; Ochman and Wilson 1986). Hori and Osawa cited Hori’s (1976) data for 5S rRNA, which was used to produce a phylogeny of eukaryotes and prokaryotes. Assuming that the ancestors of man and Xenopus diverged 300 Myr ago, they estimated $t_{ES}$ to be 37 ± 26 Myr ago. This assumes that 5S rRNA evolves at the same rate in vertebrate and enterobacterial lineages. Ochman and Wilson (1986), using similarity values ($S_{ab}$) for 16S rRNA comparisons, estimated $t_{ES}$ to be 110–150 Myr ago. The average number of substitutions per synonymous site between E. coli and S. typhimurium from the 23 genes examined here is 0.90, and the rate of synonymous substitutions per year ($V_S$) can be derived for the above estimates of $t_{ES}$. If $t_{ES} = 37$ Myr, then $V_S = 12.4 \times 10^{-9}$ substitutions/site/year. This is ~2–3 times higher than the
average rate \((4.7 \times 10^{-9})\) for 35 mammalian genes (Li et al. 1985b). If \(t_{ES} = 130\) Myr ago, then \(V_S = 3.5 \times 10^{-9}\), which is somewhat lower than the mean value for mammalian genes. These rate estimates are perhaps surprising. Since bacteria have a short generation time, their number of DNA replications per unit time could be much greater than that of mammals. Then, unless mammals have a comparatively inefficient DNA repair system, the rate of mutation should be much higher in bacteria. In turn, \(V_S\) should also be considerably higher in bacteria. Therefore, we conclude either (1) that the estimates of \(t_{ES}\) cited above are too high or (2) that selection against synonymous mutations is more effective in these bacteria than in mammals, perhaps owing to a much larger population size or stronger selective constraints on synonymous-codon usage (Li, accepted).

The present study indicates that \(V_S\) in Enterobacteria is influenced by natural selection differentiating between alternative synonymous codons. This is also likely to be true in multicellular organisms, although in the latter the determinants of synonymous-codon usage remain to be elucidated. Selection on synonymous codons is perhaps the most subtle form of natural selection detected; yet, because of the small selective differences involved, it is probably the most pervasive. Since the \(V_S\) apparently can vary widely among genes, separate molecular clocks of synonymous substitution should be used for genes with different levels of synonymous-codon usage bias.

Acknowledgments

We are grateful to Ronald C. Wek, Shahla Thompson, and Peter McWilliam for access to their unpublished DNA sequence data and to Robert J. Schwartz for plotting the figure. This work was supported by National Institutes of Health grant GM 30998.

LITERATURE CITED


WALTER H. FITCH, reviewing editor

Received August 5, 1986; revision received November 20, 1986.
The nucleotide sequences corresponding to bovine $\alpha_{s2}$- and $\beta$-casein mRNAs have been determined by cDNA analysis. Both sequences appear to be complete at their 5’ ends. The nucleotide sequence of $\alpha_{s2}$-casein, when compared with the corresponding casein A sequence, helps to define the boundaries of a large amino acid repeat (∼80 residues) whereas comparisons with the nucleotide sequences of rat $\gamma$- and mouse e-casein mRNAs also reveal extensive sequence similarities. An alignment of these four sequences shows that the divergence of their translated regions has been characterized by the duplication and deletion of discrete segments of sequence that probably correspond to exons. A high degree of nucleotide substitution is also found when the four sequences are compared, except for well-conserved leader-peptide and phosphorylation-site sequences and, to a lesser extent, the 5’-untranslated regions. Similar comparison of the bovine and rat $\beta$-caseins shows that their divergence has involved a high rate of nucleotide substitution but that no major insertions or deletions of sequence have occurred. The several splice sites that have been defined in the rat $\beta$-casein gene are likely to have been conserved in the bovine. The contrasting evolutionary histories of the $\alpha$- and $\beta$-casein coding sequences correlate with the distinctive functions of these proteins in the casein micelle system in milk.

Introduction

The caseins comprise the major protein fraction of the milk from most species. Their function is to transport calcium phosphate in milk—and hence to provide the suckling infant with a source of calcium and phosphorus for bone formation—as well as to contribute to the requirement for amino acids.

Bovine milk contains four caseins whose genes are linked in the order $\alpha_{s1}$-, $\alpha_{s2}$-, $\beta$-, and $\kappa$-casein (Grosclaude et al. 1973, 1979), and these are expressed in a coordinate manner during lactation. The caseins have been extensively studied at the protein
level, and their amino acid sequences are known (Mercier et al. 1971, 1973; Ribadeau-Dumas et al. 1972; Brignon et al. 1977).

The two α-caseins and β-casein are described as "calcium-sensitive" because they precipitate in the presence of low concentrations of this cation (Waugh 1971). In milk, the caseins occur as very large aggregates termed casein micelles, in which the calcium-sensitive caseins are maintained in stable suspension as a result of their interaction with κ-casein (Mackinlay and Wake 1971). Approximately 5% of the weight of casein micelles consists of colloidal calcium phosphate, which is sequestered by the micelles through interaction of the inorganic component with clustered serine phosphate residues of the calcium-sensitive caseins (Sleigh et al. 1979). In this way calcium phosphate is prevented from precipitating out of milk and is maintained in a form that can be readily assimilated by the young animal.

Phosphorylation of the serine clusters occurs posttranslationally as the result of the activity of one or more specific casein kinases located in the Golgi complex (Mackinlay et al. 1977). The recognition sequence for the casein kinase(s) is Ser/Thr-X-Y, where X may be any amino acid and Y is a glutamate, aspartate, or serine phosphate residue (Mercier 1981). The phosphorylation sites typically occur in the sequence Ser-Ser-Glu-Glu, sometimes preceded by Ser-X or extended in the other direction by Ser-X-Glu. It has been shown for the three rat calcium-sensitive casein genes that a splice junction occurs between the pair of glutamate codons so that sequences from two exons are involved in creating the clustered phosphorylation sites (Jones et al. 1985; Yu-Lee et al. 1986).

The caseins, being secreted proteins, possess N-terminal signal peptides that direct the passage of the newly synthesized polypeptides into the lumen of the endoplasmic reticulum, whereupon the signal peptides are removed from the primary translation products to yield the mature caseins. The leader peptide sequences are very similar among all calcium-sensitive caseins examined. This was first observed at the amino acid level for the ovine caseins (Mercier et al. 1978) and has led to the proposition that the calcium-sensitive caseins are evolutionarily related. The sequences of cDNAs determined subsequently for the caseins of the rat (Rattus rattus) (Blackburn et al. 1982; Hobbs and Rosen 1982), mouse (Mus musculus) (Hennighausen et al. 1982), guinea pig (Cavia cobaya) (Hall et al. 1984a, 1984b), and cow (Bos taurus) (Stewart et al. 1984) support this notion. The calcium-sensitive caseins are therefore thought to constitute a gene family that has arisen from an ancestral sequence by an extensive series of duplications and modifications while retaining similar 5'-untranslated, leader-peptide, and phosphorylation-site sequences. The remainder of the sequences appear to be subject to few functional constraints and have accumulated many nucleotide substitutions, as revealed by comparative sequence analysis.

In the present paper we present the complete cDNA sequences for bovine αS2- and β-caseins, thereby completing our characterization of the bovine caseins at the mRNA level. The relationship of bovine αS2-casein to cavine casein A, rat γ-, and murine ε-caseins is described herein and shows that these αS2-like caseins have diverged through extensive sequence rearrangements and nucleotide substitutions. The divergence of the bovine and rat β-caseins, by contrast, appears to have only involved extensive nucleotide substitution.

**Material and Methods**

The construction and identification of pBβC14 has been described (Willis et al. 1982). Plasmids pBβC16 and pBαS2C23 were isolated from the same cDNA library.
Because these three clones were incomplete copies of their corresponding mRNAs, another cDNA library was constructed using the method of Land et al. (1981). This cDNA library was screened by colony hybridization (Grunstein and Wallis 1979), using [α-32P]dATP-labeled *Hinf*I restriction fragments from pBB468 and pBas2C23. Plasmids pBB468, pBas2C170, and pBas2C411 were identified.

All sequencing was performed by using the method of Maxam and Gilbert (1980). Free-energy values for RNA secondary structure were calculated using the method of Ninio (1979) as detailed in de Wachter et al. (1982).

Computer alignments (figs. 2C, 4) were obtained by using the method of Taylor (1984) as implemented in the align program in the package described by Harr et al. (1986). The parameters were set to count 1 for each mismatch plus 1 for each gap introduced plus 1 for each residue that has a gap opposite it (i.e., gap penalty = 1 + length of gap).

**Results and Discussion**

**cDNA Sequences**

Restriction maps and sequencing strategies for the αs2- and β-casein cDNA clones are shown in figure 1.

Plasmid pBas2C170 contained an insert of 669 bp, excluding the GC tails. Plasmid pBas2C411 contained an insert of 515 bp, including a poly A tail of 13 residues but excluding the GC tails. The overlapping sequences imply an mRNA of 1,024 nucleotides, excluding the poly A tail with 5'-untranslated, translated, and 3'-untranslated regions of 56, 666, and 302 nucleotides, respectively (fig. 2). It is likely that the sequence is complete at the 5' end, since it commences with the dinucleotide AT (see below) and is directly comparable to the 5'-untranslated region determined from the analysis.

![Restriction maps of αs2- and β-casein cDNAs and strategies for their sequence determination.](image-url)

**Fig. 1.**—Restriction maps of αs2- and β-casein cDNAs and strategies for their sequence determination. Coding sequences are indicated by thickened lines; jagged lines represent GC tails; and the numbers define the orientations, where known, of the cDNA inserts with respect to the pBR322 vector (all cDNAs cloned into the *Pst*I site at position 3612). Arrows represent the extent and direction of sequence obtained from the restriction sites indicated by short vertical lines. Key: A = *AatI*; D = *DdeI*; F = *Hinf*I; H = *HaeIII*; K = *KpnI*; 9 = *Sau96 I*; and Y = *HpaII*.
Fig. 2.—Alignment of the αs2-like casein cDNAs. The sequences for bovine αs2, cavine A (Hall et al. 1984b), rat γ- (Hobbs and Rosen 1982), and mouse e-casein (Hennighausen et al. 1982) are compared, with similar sequences aligned vertically. 2A: 5'-Untranslated regions. Numbers refer to nucleotides with the last nucleotide of the untranslated region being set at −1. 2B: Translated regions. Numbers refer to codons, with the first codon of the mature proteins set at +1. U and D refer to the upstream and downstream portions, respectively, of the large duplication. Except for their C-terminal extremities, the mature protein-coding regions were aligned as follows: the closely related bovine and cavine sequences were aligned on the basis of dot-matrix comparisons (performed as described by Reisner and Bucholtz [1983]), gaps in these sequences being positioned to correspond to putative splice junctions. The more divergent rat and mouse sequences.
were fitted to this alignment in a similar manner, viz., by identifying segments of these sequences that correspond to putative individual exons (see text) and then visually aligning these to corresponding segments of the bovine and canine sequences. The alignment downstream of bovine codon 116 was arrived at by considering amino acid sequences and aligning amino acids that are either identical or similar. This alignment indicates that the last two codons were not part of the duplication, a conclusion consistent with the presence of a splice junction at this position in the p-casein gene (see fig. 4). 2C: 3’ untranslated regions. Numbers refer to nucleotides. The polyadenylation signals are underlined. The bovine and canine 3’ noncoding regions were aligned using the ALIGN program (see Material and Methods).
of the rat \(\gamma\)-casein cDNA and genomic sequences (Yu-Lee and Rosen 1983). The predicted leader-peptide sequence is identical to the published ovine \(\alpha_{s_2}\)-casein leader-peptide sequence (Mercier et al. 1978). The predicted mature protein sequence varies from the published bovine \(\alpha_{s_2}\)-casein A sequence (Brignon et al. 1977) only at amino acid 87, which is predicted to be glutamine rather than glutamate.

Plasmid pB\(\beta\)C468 contained an insert of 945 bp, excluding the GC tails. Plasmid pB\(\beta\)C16 contained an insert of 800 bp, excluding the GC tails and including a poly A tail of 19 residues. The overlapping sequences imply a mRNA of 1,089 nucleotides, excluding the poly A tail, with 5'-untranslated, translated, and 3'-untranslated regions of 56, 672, and 361 nucleotides, respectively (fig. 4). The first base of the 5'-most \(HinfI\) site shown on the restriction map is continuous with the G tail. Another \(\beta\)-casein cDNA isolated, pB\(\beta\)C319, also displays this feature. On this basis and from a comparison with rat \(\beta\)-casein cDNA and genomic sequences (Jones et al. 1985), we assume that the sequence presented is complete at the 5' end. The predicted leader-peptide sequence is identical to the published ovine \(\beta\)-casein leader-peptide sequence (Mercier et al. 1978), and the predicted mature protein sequence agrees with the published bovine \(\beta\)-casein A \(_2\) sequence (Ribadeau-Dumas et al. 1972), with the following exceptions: (1) amino acid 117, which is predicted to be glutamate rather than glutamine; (2) amino acids 137 and 138, which are predicted to be leucine-proline rather than proline-leucine; (3) amino acid 175, which is predicted to be glutamine rather than glutamate; and (4) amino acid 195, which is predicted to be glutamate rather than glutamine.

In both mRNAs the protein-coding region is predicted to commence with the first AUG encountered, and both conform to the observations of Kozak (1981). Both were found to contain the conventional polyadenylation signal (Fitzgerald and Shenk 1981).

The \(\alpha_{s_2}\)-Like Caseins

The 5'-Untranslated Regions

A comparison of the 5'-untranslated regions of the \(\alpha_{s_2}\)-like casein cDNAs is presented in figure 2A. The canine and murine sequences are incomplete at their 5' ends. Comparisons of these sequences—as well as those of rat \(\beta\)- and the three \(\alpha_{s_1}\)-like caseins for which sequences are also available—reveal a number of common features. All begin with the dinucleotide AT, which may be related to the efficiency of initiation of their translation, as in other families of mRNAs in which the first two bases are conserved (Breathnach and Chambon 1981). Their translation-initiation sites are also conserved, with a purine -3 and a C -1 to the initiating AUG.

Bases -14 to -11 consist of the sequence AGGA in the \(\alpha_{s_2}\)-like and \(\beta\)-caseins and of AGAT in the \(\alpha_{s_1}\)-like caseins. These have been shown to be the junctions between exons I and II in the three rat caseins, in bovine \(\alpha_{s_1}\)-casein (Yu-Lee et al. 1986), and in bovine \(\beta\)-casein (J. Bonsing, unpublished results).

In the \(\alpha_{s_2}\)-like caseins, the sequences from -36 to -13 contain a string rich in pyrimidines followed by a string rich in purines. These are able to form stable hairpin loops. For bovine \(\alpha_{s_2}\)-casein, the calculated \(\Delta G\) value in kilocalories per mole for this structure is -7.4 (when the thermodynamic binding model is used) or -9.6 (when the empirical binding model, as detailed in de Wachter et al. [1982], is used). This structure comprises nucleotides -36 to -27 base pairing with nucleotides -21 to -13 to form a 9-bp stem including a single G:U bp, a single looped-out nucleotide (-31), and an unpaired loop containing nucleotides -26 to -22. Similar structures are possible.
for cavine A and rat γ-casein with calculated values of \(-7.1\) or \(-9.6\) and \(-5.7\) or \(-6.6\) kcal/mol, respectively. Comparable secondary structures cannot be drawn for other published casein sequences, although they possess similar purine and pyrimidine strings. Occurring within the stem portions of these loops is the sequence GGAA, also found in the region of 18S rRNA proposed to interact with mRNAs during initiation of translation (Baralle 1983). These secondary structures may therefore modulate the rate of initiation of translation of the αS₂-like casein mRNAs, the products of which are all minor components in the milks in which they occur.

The Translated Regions

Figure 2B presents an alignment of the translated regions of the four αS₂-like casein cDNAs. A prominent feature of these sequences is a repeated structure, first detected in the amino acid sequence of bovine αS₂-casein (Brignon et al. 1977) and extended here to give a tandem repeat consisting of upstream (U) codons 33–125 and downstream (D) codons 126–205 in the bovine sequence and codons 33–113 and 114–206 in the cavine sequence. Rat codons 34–104 and 105–161 also correspond to portions of this tandem repeat. In murine ε-casein the repeat is absent, but this sequence has a unique repetition of codons 12–33 in codons 53–75.

Predominantly hydrophilic amino acid sequences are encoded by (1) the bovine sequence up to approximately codon 68, (2) the sequences in other species aligned with this section, and (3) the corresponding downstream repeated sequences. These are shown divided into a number of segments in figure 3. The boundaries of these segments, most of which are likely to correspond to existing or ancestral splice junctions, have been defined according to two criteria: (1) identity with established splice junction sequences, particularly those adjacent to phosphate clusters (Jones et al. 1985; Yu-Lee et al. 1986), with such boundaries including those at codons 11–12, 41–42, 59–60, and 67–68, and (2) coincidence with gaps in the alignment, e.g., following codon 33, where gaps occur at this position in the mouse and rat sequences and following codons 67–68, where a gap occurs at this position in the cavine sequence. A boundary following codon 50 may be inferred since a gap occurs after codon 141, the corresponding position in the upstream repeat. Further evidence for boundaries at codons 50–51 and 59–60 is provided by the deletion that characterizes the αS₂-casein D variant

![Fig. 3.—Summary of the sequence alignments of the αS₂-like caseins. Numbering refers to the first codon of each segment, except for termination codons and cavine 67, murine 33 and 75, and rat 33 and 104, which are the last codons in the sections of sequence that precede them. 5'ut = 5'-Untranslated; L = leader peptide-coding region; P = major phosphorylation site found within this segment; 3'ut = 3'-untranslated region; and Aₙ = poly (A) tail. Minor gaps in the sequence alignment of fig. 2 have been omitted. Also shown are regions of 3'-untranslated sequences that show detectable similarity on the basis of dot-matrix analysis.](image-url)
(Grosclaude et al. 1979). Although the presence of splice sites at these positions remains to be experimentally demonstrated, the situation clearly parallels that in the rat β-casein gene, where the hydrophilic phosphate cluster–containing N-terminus of the molecule has been shown to be encoded by a series of short exons postulated to have arisen as the result of a series of duplications of an ancestral sequence (Jones et al. 1985).

The bovine sequences from codons 68–125, the C-terminal codons 150–207, and the corresponding sequences in the other αs2-like casein encode predominantly hydrophobic protein domains. These sequences have been subject to a high rate of amino acid replacement. For example, of the 42 nucleotide differences found between bovine codons 83–125 and cavine codons 68–113, 30.5 (73%) are replacement differences. We believe that this implicates selective neutrality toward hydrophobic amino acid replacements in these domains—and that they do not code for structurally important amino acid sequences, in contrast to the well-conserved phosphorylation-site sequences. However, constraints have acted to maintain a relatively high degree of hydrophobicity in these domains as a whole, a characteristic important for the protein-protein interactions involved in casein micelle formation (Waugh 1971). Figure 3 summarizes the relationships between the four αs2-like caseins.

The β-Caseins

An alignment of the bovine and rat β-casein cDNAs is presented in figure 4. Many nucleotide substitutions have occurred during the divergence of these caseins, but no major insertions, deletions, or other sequence rearrangements—such as are found in the αs2- and αs1-like caseins (Stewart et al. 1984)—are evident. From the amino acid sequence of human β-casein (Greenberg et al. 1984), we infer that this is also the case for that species. In the rat, four short exons encode the N-terminal hydrophilic sequence that contains a single phosphate cluster, whereas the remainder of the molecule, of markedly hydrophobic character, is encoded by a single exon. The conservation of rat and bovine sequences that span the splice sites of the rat β-casein gene (indicated by arrowheads in fig. 4) suggests that the arrangement of exons and introns has been conserved in the rat and bovine β-casein genes.

The absence of major sequence rearrangements in the β-caseins and the maintenance of hydrophilic N-terminal and hydrophobic C-terminal domains, despite many amino acid replacements, points to the operation of functional constraints acting to conserve the overall architecture of the molecule. Consistent with this conclusion is recently obtained evidence that shows β-casein to be important in determining the surface properties of casein micelles and essential for curd formation when milk is clotted by the proteolytic enzyme chymosin (Pearse et al. 1986). Curd formation is physiologically important since it ensures the retention of milk protein in the stomach of the infant and allows further digestion to occur.

A similar conservation of structure is found for the κ-caseins, which are also important in determining properties of casein micelles, including the maintenance of micelle stability and the initiation of curd formation when this protein is cleaved by chymosin at a specific Phe-Met bond. The importance of the β- and κ-caseins is indicated by their occurrence in all eutherian milks so far examined (Jenness 1979).

The α-caseins, by contrast, occur in varying amounts or may be absent altogether, as is the case for human milk. These caseins appear to be much less important for curd formation than the β- and κ-caseins (Pearse et al. 1986), and their role is probably restricted to determining the capacity of casein micelles for calcium phosphate transport.
BOVINE 1 ATTC AGCTCCTTCTACCTCTTCCTCTGCTCTTTGAAAAGTAAATGAGGAGGAGC
RAT 1 ATCCCTGAGCTACTTCTTCCTTTGCTCTTGGGGGAGCAGGCGCCATCT

BOVINE 57 AGTGAAGCTCTTCCCTTGCTTCAAGGCTTGCTCTGGGCTGGTGC
RAT 53 GATGAAGCTCTTCCCTTGCTTCAAGGCTTGCTCTGGGCTGGTGC

BOVINE 118 TACATGACCTGGCTCTGGCTGCAAGGCTTGCTCTGGGCTGGTGC
RAT 114 TACATGACCTGGCTCTGGCTGCAAGGCTTGCTCTGGGCTGGTGC

FIG. 4.—The nucleotide sequence of bovine β-casein cDNA shown aligned with the corresponding rat sequence (Blackburn et al. 1982). Alignment was carried out using the ALIGN program (see Material and Methods). Translation-initiation and -termination codons and the first codons of each of the mature proteins are underlined. Arrowheads indicate the position of splice junctions in the rat sequence (Jones et al. 1985).

Thus, the structure-function relationships of the different caseins are reflected in the manner in which each has been modified during the evolution of mammalian species.
Acknowledgments

We wish to thank Dr. A. Reisner and G. Kornfeld for assistance with computer analyses. This work was supported by the Australian Research Grants Scheme, the Australian Dairy Research Committee, and by Commonwealth Postgraduate Awards to A.F.S. and J.B.

LITERATURE CITED


WALTER M. FITCH, reviewing editor

Received July 7, 1986; revision received November 14, 1986.
DNA-DNA Hybridization Evidence of the Rapid Rate of Muroid Rodent DNA Evolution

François M. Catzeflis,* Frederick H. Sheldon,† Jon E. Ahlquist,‡ and Charles G. Sibley†

*Institut des Sciences de L'Évolution, Université Montpellier II (U.S.T.L.);
†Tiburon Center, San Francisco State University; and ‡Department of Zoological and Biomedical Sciences, Ohio University

Single-copy nuclear DNAs (scnDNAs) of eight species of arvicoline and six species of murine rodents were compared using DNA-DNA hybridization. The branching pattern derived from the DNA comparisons is congruent with the fossil evidence and supported by comparative biochemical, chromosomal, and morphological studies. The recently improved fossil record for these lineages provides seven approximate divergence dates, which were used to calibrate the DNA-hybridization data. The average rate of scnDNA divergence was estimated as 2.5%/Myr. This is ~10 times the rate in the hominoid primates. These results agree with previous reports of accelerated DNA evolution in muroid rodents and extend the DNA-DNA hybridization data set of Brownell.

Introduction

The reconstruction of phylogenies and the determination of the rate—or rates—of genomic evolution are among the most controversial facets of molecular evolution. Herein, we present DNA-DNA hybridization evidence of the branching pattern of the lineages of seven genera and 14 species in the rodent taxa Microtinae and Murinae (superfamily Muroidea), as defined by Carleton and Musser (1984). Thanks to an improved fossil record, we also present evidence that the average rate of genomic evolution in the muroid rodents is much faster than the average rates in hominoids and birds.

Our DNA-hybridization data, as well as those of paleontological (Repenning 1968; Chaline 1974, 1980), morphological (Hooper and Musser 1964; Carleton 1981), chromosomal (Gamperl 1982; Koop et al. 1984; Modi 1987), and other biochemical (Graf 1982; Bonhomme et al. 1985) studies, suggest that, within the Microtinae (=Arvicolidae of Chaline [1974, 1980]), (1) Palearctic and Nearctic species of voles (Microtus) are more closely related to one another than to species of water voles (Arvicola) or red-backed voles (Clethrionomys), (2) Microtus and Arvicola (tribe Microtini) are more closely related to one another than either is to Clethrionomys (tribe Clethrionomyini), (3) the tribe Lemmini (the lemmings, Lemmus, and relatives) is the sister group of the Microtini-Clethrionomyini clade, and (4) that, within the Murinae (=Muridae of Petter [1966] and Eisenberg [1981]), the rats (Rattus) are the sister taxon of the clade comprising mice (Mus) and wood mice (Apodemus), the biochem-

1. Key words: rodent phylogeny, rate of DNA evolution, DNA-DNA hybridization.

Address for correspondence and reprints: Dr. François M. Catzeflis, Institut des Sciences de L’Évolution, Université Montpellier II (U.S.T.L.), Place Eugène Bataillon, F-34060 Montpellier, France; or Dr. Charles G. Sibley, Tiburon Center, San Francisco State University, Box 855, Tiburon, California 94920.

© 1987 by The University of Chicago. All rights reserved.
0737-4038/87/0403-0004$02.00

242
ical, chromosomal, and morphological differences among these latter three genera being much greater than those between the voles and lemmings.

In addition, the common ancestor of the Microtinae (hereafter referred to by the common name arvicoline) and Murinae is unknown (Petter 1966; Chaline et al. 1977; Carleton and Musser 1984).

Several molecular studies have indicated that rodent DNAs evolve much faster than those of large mammals. The first DNA-DNA hybridization study suggesting this rate difference was that of Laird et al. (1969). They observed a greater genetic distance between mouse and rat than between cow and pig, although available fossil evidence suggested that cow and pig are genealogically more distinct than mouse and rat. Others to observe an apparent rapid rate of rodent DNA evolution include Rice (1971), Benveniste et al. (1977), Brownell (1983), and Wu and Li (1985). Britten (1986) reviewed the evidence for differences in rates in different taxonomic groups and came to the conclusion that rodent DNA has evolved approximately five times as fast as that of hominoids and birds.

In her study, which is the most complete set of rodent DNA-hybridization comparisons to date, Brownell (1983) emphasized that the poor fossil record introduces substantial error into the calibration of the amount of genomic change that occurs per unit of time. Sarich (1972), Sarich and Cronin (1977), and Wilson et al. (1977), arguing for overall constancy in DNA evolutionary rates among groups, also made this point and reasoned that apparent discrepancies in rates of DNA evolution are primarily artifacts of a poor fossil record.

However, recent work on the fossils of arvicoline (Chaline 1974, 1980, 1986) and murines (Flynn et al. 1985; Jaeger et al. 1985, 1986) has increased our knowledge of the history of these groups. Together, the molecular data on degrees of genetic divergence and this improved evidence of divergence dates permit a more accurate estimate of the rate of DNA evolution—and hence a determination of the accuracy of the molecular clock.

Material and Methods

Descriptions of our methods have been published by Sibley and Ahlquist (1981, 1983, 1984).

In brief, DNA extracts were obtained from the nuclei of ethanol-preserved tissue cells, purified, and sheared by sonication into fragments with an average length of 500 bases. Single-stranded fragments of the species to be used as radiolabeled tracers were reassocated to Cot 1,000 at 50 C in 0.48 M sodium phosphate buffer, and repeated sequences were removed by hydroxyapatite (HAP) chromatography. The single-copy nuclear DNA (scnDNA), representing 50%-60% of the total genome by volume and ≥99% by complexity, was labeled with 125I. DNA-DNA hybrids were formed from a mixture composed of one or two parts (=200 or 400 ng) of tracer DNA and 1,000 parts of sheared, whole DNA of the driver species. These proportions ensured that only ~1%-2% of the duplexes formed would be between tracer fragments. The incipient hybrids were denatured at 100 C, then incubated to Cot 16,000 at 60 C in 0.48 M phosphate buffer to permit the single strands to form hybrid duplexes.

After incubation, the buffer was diluted to 0.12 M and the hybrids were bound to HAP columns immersed in a temperature-controlled water bath at 55 C. The temperature was then raised in 2.5-degree C increments from 55 to 95 C. At each of 17
temperatures, the single-stranded fragments produced by the melting of duplexes were eluted with 0.12 M sodium phosphate buffer. The radioactivity of each sample was counted, and the data were used to calculate $T_{50}^R$ values.

$T_{50}^R$ is the temperature at which 50% of all potential hybrid DNA sequences retain their duplex form and 50% have dissociated into single-strands. Delta $T_{50}^R$ is the difference between the $T_{50}^R$ value of a homoduplex control (formed by tracer and driver DNAs derived from the same individual) and any heteroduplex hybrid (formed by tracer and driver DNAs of different individuals) measured in the same experiment. Delta $T_{50}^R$ measures the median sequence divergence between the genomes of two taxa. It is an estimate of their average percent nucleotide difference, based on the finding that an $\sim 1.0$–degree C reduction in melting temperature corresponds to a 1% difference in nucleotide sequence of the DNAs being compared (Britten et al. 1974). As do $T_mR$ (Benveniste et al. 1977; O'Brien et al. 1985) and TMH (Koop et al. 1986), $T_{50}^R$ takes into account the final percent hybridization.

The scientific and common names of the species compared in this study are listed in table 1. Of these species, four were radiolabeled: bank vole (*Clethrionomys glareolus*), sibling vole (*Microtus epiroticus*), western Mediterranean short-tailed mouse (*Mus spretus*), and common rat (*Rattus norvegicus*).

**Results and Discussion**

**Reciprocity**

A matrix of the delta $T_{50}^R$ distances derived for pairs of rodents in this study is presented in table 2. From these data, the degree of reciprocity—i.e., the degree to which the distance from labeled taxon A to driver taxon B agrees with the distance from labeled B to driver A—was calculated. The average delta $T_{50}^R$ value for comparisons between labeled *Microtus* DNA and driver *Clethrionomys* DNA equals 10.02 ± 0.71 ($n = 10$). The reciprocal distance is 11.02 ± 0.85 ($n = 5$). *Mus to Rattus* and

---

**Table 1**

<table>
<thead>
<tr>
<th>Family and Common Name (Genus Species)</th>
<th>Locality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Microtinae:</strong></td>
<td></td>
</tr>
<tr>
<td>Sibling vole (<em>Microtus epiroticus</em>)</td>
<td>Bulgaria</td>
</tr>
<tr>
<td>Meadow vole (<em>Microtus pennsylvanicus</em>)</td>
<td>Connecticut</td>
</tr>
<tr>
<td>Prairie vole (<em>Microtus ochrogaster</em>)</td>
<td>Kansas</td>
</tr>
<tr>
<td>Water vole (<em>Arvicola terrestris</em>)</td>
<td>Switzerland</td>
</tr>
<tr>
<td>Bank vole (<em>Clethrionomys glareolus</em>)*</td>
<td>Switzerland</td>
</tr>
<tr>
<td>Northern red-backed vole (<em>Clethrionomys rutilus</em>)</td>
<td>Alaska</td>
</tr>
<tr>
<td>Boreal red-backed vole (<em>Clethrionomys gapperi</em>)</td>
<td>Connecticut</td>
</tr>
<tr>
<td>Siberian lemming (<em>Lemmus sibiricus</em>)</td>
<td>Alaska</td>
</tr>
<tr>
<td><strong>Murinae:</strong></td>
<td></td>
</tr>
<tr>
<td>House mouse (<em>Mus musculus</em>)</td>
<td>Laboratory strain</td>
</tr>
<tr>
<td>Western Mediterranean short-tailed mouse (<em>Mus spretus</em>)*</td>
<td>France</td>
</tr>
<tr>
<td>Common rat (<em>Rattus norvegicus</em>)*</td>
<td>Laboratory strain</td>
</tr>
<tr>
<td>Wood mouse (<em>Apodemus sylvaticus</em>)</td>
<td>Switzerland</td>
</tr>
<tr>
<td>Striped field mouse (<em>Apodemus agrarius</em>)</td>
<td>Italy</td>
</tr>
<tr>
<td>Yellow-necked mouse (<em>Apodemus flavicollis</em>)</td>
<td>Switzerland</td>
</tr>
</tbody>
</table>

* Radio-labeled taxon.
Table 2
Mean ± SD Delta T$_{so}$H Distances between Pairs of Muroid Rodents

<table>
<thead>
<tr>
<th></th>
<th>Microtus epiroticus</th>
<th>Clethrionomys glareolus</th>
<th>Mus spretus</th>
<th>Rattus norvegicus</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. epiroticus</td>
<td>4.3 ± 0.1 (4)</td>
<td>10.4 ± 0.4 (3)</td>
<td>32.4 ± 0.5 (4)</td>
<td>ND</td>
</tr>
<tr>
<td>M. pennsylvanicus</td>
<td>4.5 ± 0.2 (3)</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>M. ochrogaster</td>
<td>8.7 ± 0.3 (5)</td>
<td>10.8 ± 0.5 (4)</td>
<td>30.8 ± 1.0 (4)</td>
<td>31.4 ± 0.9 (6)</td>
</tr>
<tr>
<td>Arvicola terrestris</td>
<td>10.2 ± 0.5 (2)</td>
<td>1.8 ± 0.3 (4)</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>C. glareolus</td>
<td>10.1 ± 1.0 (5)</td>
<td>2.9 ± 0.7 (4)</td>
<td>31.7 ± 2.1 (2)</td>
<td>ND</td>
</tr>
<tr>
<td>C. gapperi</td>
<td>9.8 ± 0.2 (3)</td>
<td>12.6 ± 0.5 (4)</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>L. sibiricus</td>
<td>30.0 ± 0.8 (2)</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>M. spretus</td>
<td>31.4 ± 1.0 (4)</td>
<td>32.6 (1)</td>
<td>20.1 ± 0.8 (6)</td>
<td>ND</td>
</tr>
<tr>
<td>A. sylvaticus, A. agrarius, and A. flavicollis.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE.—Labeled species are given on the horizontal axis. Numbers in parentheses are number of comparisons. ND = not determined.

vice versa are 19.94 ± 0.76 (n = 23) and 20.15 ± 0.79 (n = 6), respectively, and Microtus to Mus and vice versa are 30.93 ± 1.08 (n = 6) and 32.40 ± 0.49 (n = 4), respectively. When t-tested, these reciprocal values indicate no significant difference at the P = 0.05 level.

Brownell (1983) reported a significant degree of nonreciprocity between delta mode values of Mus and Rattus. However, we found that when delta T$_{so}$H values were calculated from data provided by Dr. Brownell, these values exhibited excellent reciprocity: Mus to Rattus, 20.7 (n = 2); Rattus to Mus, 20.0 ± 1.7 (n = 3).

The mean percent reciprocal deviation was also calculated, using the formula of Champion et al. (1974). For Microtus to Clethrionomys, the mean deviation from the average reciprocal distance for 30 tests was 3.6%. In Mus-to-Rattus comparisons, 120 tests yielded a 2.5% mean deviation. Even in tests between arvicolines and murines, the mean deviation is small: 2.4% for 24 Mus-to-Microtus tests. These deviations are much lower than the values commonly observed for immunological distances derived on the basis of data for serum albumins and transferrins (e.g., such values are 6%–8% in Champion et al. [1974]).

Complete reciprocity, demonstrated even for taxa as distant as Mus and Microtus, suggests that all reciprocal values are likely to be equivalent, including those not actually measured. Making this assumption, we averaged all reciprocal distances (i.e., folded the distance matrix) before clustering our data into phylograms (see below).

Relative-Rate Tests

At least four sets of comparisons qualify as relative-rate tests (i.e., those tests in which two of three taxa are more closely related to one another than either is to the third [Sarich and Wilson 1973]). These sets are outgroup Mus to ingroups Arvicol a and Microtus, Rattus to Arvicol a and Clethrionomys, Microtus to Mus and Apodemus, and Rattus to Apodemus and Mus. From the data in table 2, it is apparent that the distances from the outgroup to each of the two ingroups are essentially equal (t-test,
Thus, in all cases in which the outgroup is unambiguous, our data indicate that the scnDNAs of arvicoline and murine rodents are evolving at approximately the same average rate. Similarly, Brownell (1983) found that nucleotide substitution was occurring at a uniform average rate in several arvicoline and sigmodontine rodent lineages.

Comparisons with Previously Published Data

To compare our delta T50H values with the distances derived in other DNA-hybridization studies, we calculated the relationships among mode, Tm, and T50H, using distances ranging from delta 4 to delta 34 (data to be published elsewhere). Delta mode and delta T50H are related by the following power regression:

\[ T_{50H} = 0.8 \text{ Mode}^{1.17} \]

\[(n = 84; r = 0.996).\]

This equation was used to estimate delta T50H from modal values published by Brownell (1983). The relationship between delta Tm and delta T50H is described by the following function:

\[ T_{50H} = 0.63 T_m^{1.32} \]

\[(n = 43; r = 0.993).\]

This regression was used to estimate T50H from the Tm values of Laird et al. (1969), Rice and Straus (1973), and Rice (1974).

Our estimates of the mean number of differences per site conform well to some of the previously published distances between the same taxa. Brownell’s (1983) delta mode of 8.6 ± 0.6 between Microtus and Clethrionomys corresponds to a delta T50H of 9.9, a value nearly identical to our 10.1 ± 0.7 (n = 13). For Mus to Rattus, Laird et al. (1969) and Rice and Straus (1973) derived delta Tm values of 14 and 14.9, respectively. These distances correspond to delta T50H’s of 20.4 and 22.1 and are therefore close to our average distance 19.9 ± 0.8 (n = 23) and to that which we calculated from Brownell’s raw data (20.2). The distances between the Microtinae and Murinae have been estimated from Rice (1974) and Brownell (1983) to range from delta T50H 28.2 to T50H 37.2, thus agreeing with our values of 29.4–33.2 (table 2).

The smaller delta T50H distances measured between Mus and Rattus by Benveniste et al. (1977) are possibly the result of differences in DNA-hybridizing techniques. These authors treated their samples with S1 nuclease before thermal fractionation, thus eliminating single-stranded tails, which we include in our measurements.

Phylogeny

The existence of a single average rate of scnDNA evolution among the arvicoline and murine rodents permits the use of phenetic methods for clustering taxa into phylograms. We chose the unweighted pair group method of analysis (UPGMA; Sneath and Sokal 1973), rather than least squares or other fitting algorithms, because average linkage allows the combining of matrix cells when pairwise comparisons are missing, as was the case in our data set.

To produce more accurate and additive distances, we followed the example of Koop et al. (1986) and converted the delta T50H values of table 2, which are the mean number of base-pair differences per 100 sites between the DNA sequences of two taxa, into TMH-C values (table 3). TMH-C is the mean number of substitutions per 100
Table 3
Mean + SD Delta $T_{90}$H Values, Mean Delta $T_{90}$H Values Transformed into TMH-C Distances by Correcting for Multiple Substitutions, and Estimated Times since Divergence of Various Arvicoline and Murine Clades

<table>
<thead>
<tr>
<th>Clade Pairs</th>
<th>Delta $T_{90}$ (MYBP)</th>
<th>TMH-C (MYBP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paleartic/Nearctic <em>Microtus</em></td>
<td>4.4 ± 0.2 (7)</td>
<td>4.5</td>
</tr>
<tr>
<td><em>Microtus</em>/Arvicola</td>
<td>8.7 ± 0.3 (5)</td>
<td>9.2</td>
</tr>
<tr>
<td><em>Microtini</em>/Clethrionomyini</td>
<td>10.3 ± 0.7 (17)</td>
<td>11.1</td>
</tr>
<tr>
<td><em>Microtini</em>/Clethrionomyini/Lemmini</td>
<td>11.8 ± 0.8 (10)</td>
<td>12.9</td>
</tr>
<tr>
<td><em>Mus</em>/Apodemus</td>
<td>18.3 ± 0.8 (9)</td>
<td>21.1</td>
</tr>
<tr>
<td><em>Mus</em>/Apodemus/Rattus</td>
<td>20.5 ± 0.7 (29)</td>
<td>24.1</td>
</tr>
<tr>
<td><em>Murinae</em>/Arvicolinae</td>
<td>31.4 ± 1.0 (34)</td>
<td>40.7</td>
</tr>
</tbody>
</table>

NOTE.—Numbers in parentheses are number of comparisons.

*Calculated using the Jukes and Cantor (1969) correction factor for multiple hits at single sites. The mean number of substitutions for the Arvicolinae/Murinae comparison is probably larger than indicated, since the Jukes and Cantor formula tends to underestimate the number of multiple substitutions as delta $T_{90}$H becomes large (Tajima 1985).

b Based on the fossil record (see text).

sites that have occurred since two taxa diverged from a common ancestor. It is calculated with Jukes and Cantor’s (1969) conversion factor for multiple substitutions (hits) per site. The phylogram of figure 1 was drawn using the TMH-C values from table 3.

![Phylogram of arvicoline and murine relationships. The branching pattern and distances were computed by average linkage using the TMH-C values from table 3.](image-url)
The subdivision of the microtines into three tribes—Microtini, Clethrionomyini, and Lemmini—has been proposed by Gromov and Poliakov (1977) on morphological grounds. Our data support such an arrangement and show that Lemmini is the sister group of the clade comprising Microtini and Clethrionomyini. Evidence from dental morphology (Chaline 1974, 1980), electrophoretic comparisons of proteins encoded by 19–22 loci (Graf 1982), and chromosomal analyses (Modi 1987) also indicate this sister-group relationship. However, comparative anatomy of the glans penis (Hooper and Musser 1964) does not.

Within the Clethrionomyini, our data indicate a slightly closer affinity between *C. glareolus* and *C. gapperi* than between either of them and *C. rutilis*. However, the G- and C-banded karyotypes of *C. glareolus* and *C. rutilis* are more similar between these two species than any of them are to those in *C. gapperi* (Modi and Gamperl 1986).

Our comparisons also suggest that *Mus* and *Apodemus* form the sister group of *Rattus*. Jacobs (1978) came to the same conclusion on the basis of fossil evidence, as did Sarich (1985) on the basis of albumin immunological data. However, on the basis of a comparative study of molar teeth, Misonne (1969) proposed a closer relationship between *Mus* and *Rattus* than between *Mus* and *Apodemus*.

*Mus spretus* and *M. musculus* are as genetically distant from one another as are species of *Microtus* or *Clethrionomyys*: all of these congeneric distances are between delta T<sub>50</sub>H 2 and delta T<sub>50</sub>H 5. Our findings also agree with the electrophoretic data (based on 24 structural loci), which show that *M. musculus* and *M. spretus* are more similar to each other than either is to several species of *Apodemus* (Bonhomme et al. 1985).

Dates, Rates, and Possible Causes

Zakrzewski (1985) has estimated that *Microtus* arrived in the Nearctic ~1.8 million years before the present (MYBP), and Chaline (1974) and Repenning (1980) dated the divergence of Nearctic and Palearctic *Microtus* as being 1.2–1.8 MYBP.

The morphological characters differentiating *Microtus* from *Arvicola* appear in fossils dating from ~2.0 MYBP (Gromov and Poliakov 1977), but it is possible to trace the lineage leading to *Arvicola* back to 3.5 MYBP (Chaline 1986, and personal communication).

Gromov and Poliakov (1977) and Chaline (1974, 1977) estimated that the Microtini and Clethrionomyini diverged ~3.7 MYBP.

*Lemmus* was distinct from *Synaptomys*, another member of the Lemmini, at 2.5–3.0 MYBP (Gromov and Poliakov 1977; J. Chaline, personal communication), and the ancestors of the Lemmini are represented by fossils that are distinct from the Microtini and Clethrionomyini at ~4.5-5.0 MYBP (Chaline 1974, 1977; Gromov and Poliakov 1977).

Recently discovered murid fossils from Pakistan and a reexamination of the oldest known murids (Jacobs 1978; Flynn et al. 1985; Jaeger et al. 1986) suggest that the *Mus-Rattus* split occurred ~8–11 MYBP. L. J. Flynn (personal communication) believes that the divergence date is probably close to 11 MYBP. The *Mus-Apodemus* divergence is dated by Flynn (personal communication) at 7–10 MYBP—i.e., within 1–2 Myr after the *Rattus* lineage branched. A succession of closely spaced divergence dates is also indicated by our DNA data (fig. 1).

The murine lineage split from the microtine lineage ≥20 MYBP (Jaeger et al.
1985; L. J. Flynn, personal communication) and more likely 25–30 MYBP (Lindsay 1978).

From these dates (summarized in table 3 and illustrated in fig. 2) and the TMH-C distances, we calculated the number of nucleotide changes that have occurred per unit of time for the muroid lineages. The average rate of DNA divergence is delta TMH-C 1.0 = 0.38 ± 0.07 Myr (n = 10). Assuming that delta 1.0 corresponds to a 1%-bp mismatch (Britten et al. 1974), 1 Myr of divergence between two species results in an ~2.5% difference in their DNAs. A rate of 2.5% nucleotide change/Myr is ~10 times as fast as the rate calculated for the hominoid primates (Sibley and Ahlquist 1984). This difference is illustrated in figure 2.

To determine the rate for hominoids, Sibley and Ahlquist (1984) used the divergence date of the Orangutan clade, set at ~13–16 MYBP by Pilbeam (1983). This gave an average rate of 0.23%/Myr. To calculate the rate for ratite birds, Sibley and

---

**FIG. 2.**—Graphic representation of single-copy genomic-rate differences between rodents (upper solid line) and hominoids (lower solid line). The slope of the rodent line is 2.5% nucleotide substitutions/Myr. The dotted lines on either side of the rodent line represent ±1 SD borders to the rodent line. The slope of the hominoid line is 0.23% nucleotide differences/Myr (from Sibley and Ahlquist 1984). The dots date events given in table 3, and the frames around the dots represent ±1 SD in two dimensions.
Ahlquist (1981) used the opening of the South Atlantic at ~80 MYBP, setting the rate at ~0.22%/Myr. Even if these fossil and vicariance dates are off by a large amount (e.g., by a factor of two in birds, as suggested by Helm-Bychowski and Wilson [1986]), which we do not believe them to be, the average genomic rate of change in rodents must have been much faster than those of hominoids and some birds.

Laird et al. (1969), Kohne (1970a, 1970b), and Kohne et al. (1970), using a Mus-Rattus divergence date of 10 Myr, found that rodents have diverged at a rate of ~2%/Myr. Benveniste et al. (1977, p. 859) analyzed thermal stability profiles of DNA duplexes of several taxa of primates and rodents and observed a “6- to 10-fold increase in the accumulation of base pair mutations in rodent cellular DNA as a function of time.” Wu and Li (1985) have discovered that rates of evolution in 11 rodent genes were faster than those of their homologues in hominoids. However, some questions have been raised concerning their use of the relative-rate test to determine these rate differences (Easteal 1985); Wu and Li used the bovine lineage as outgroup, and the ungulate-primate-rodent branching pattern is disputed. However, even if the precise branching pattern is not known, Wu and Li’s conclusions are still valid, because ungulates, primates, and rodents diverged in a short span of time, too short for the large number of nucleotide changes in the rodents to have been caused by genealogical separation rather than by differences in rates (Li and Wu 1987). Finally, as noted above, Britten (1986) concluded that the rate of DNA evolution in rodents is at least five times that in birds and hominoids.

Arguments against variable rates have usually been proposed on the basis of protein evidence (e.g., Wilson et al. 1977), but the effects of variable rates among proteins have made conclusions based on them questionable. Sibley and Ahlquist’s (see, e.g., 1983) mistaken belief in uniform rates of scnDNA evolution was based partly on the remarkable similarity of hominoid and ratite bird rates (0.23% vs. 0.22% change/Myr), partly on the constancy of rates among most passerine birds, and partly on the use of faulty relative-rate tests. The outgroup taxa employed in these rate tests (heron and plover) had slow rates of evolution, and, when employed as references to various nonpasserine ingroups, they tended to equalize discrepancies in branch lengths.

Now that the consensus of information has turned in favor of variable rates of molecular evolution, the next frontier is to discover the cause(s) of rate changes. Evidence for the rapid evolution of rodent DNA certainly points toward a generation-time effect and related phenomena (see, e.g., Laird et al. 1969; Kohne et al. 1970; Goodman 1985; Wu and Li 1985). However, Britten (1986) has suggested that generation time, population history, and selection are not likely to be the primary causes. Instead, he has proposed that the effectiveness of DNA repair mechanisms in various groups of organisms may be involved.

We conclude that there is not a single, global DNA clock ticking at the same average rate in all mammals; rather, the rate of genomic evolution in each group must be determined separately by calibrating numbers of nucleotide changes with absolute divergence dates that are derived from fossil or vicariant events. Whether the differences in average genomic rates among groups of organisms are due to differences in generation times, repair mechanisms, or other causes is not yet clear.

Acknowledgments

We thank Roy Britten, Jean Chaline, William S. Modi, and Jean-Jacques Jaeger for providing preprints of manuscripts. Marie-Claude Durette-Desset, Robert Hoff-
mann, and Fred Sibley provided tissue samples. Linda Wallace assisted in the laboratory. Funds for laboratory work were provided by Yale University and the U.S. National Science Foundation (grant BSR 83-15155 to C.G.S.). F.M.C. was supported by a postdoctoral fellowship from the Swiss National Science Foundation.

LITERATURE CITED


WAIRER M. FITCH, reviewing editor

Received March 31, 1986; revision received November 7, 1986.
Restriction-Map Variation in the Alcohol Dehydrogenase Region of *Drosophila pseudoobscura*¹

*Stephen W. Schaeffer,*⁺⁻² *Charles F. Aquadro,*⁺⁻³ *and Wyatt W. Anderson†*

*Laboratory of Genetics, National Institute of Environmental Health Sciences; and
†Department of Genetics, University of Georgia

A 32-kb region including the *Adh* structural gene was analyzed with six restriction endonucleases in 20 lines of *Drosophila pseudoobscura*, one line of *D. persimilis*, and one line of *D. miranda*. Nineteen lines of *D. pseudoobscura* from a single population were estimated to be polymorphic at one in every 15 nucleotides (*p* = 0.066). Any two homologous chromosomes chosen at random were heterozygous at one in 48 (*H* = 0.021) nucleotides. Two small insertions of 50 and 200 bp were found ~7 kb upstream from the *Adh* transcript. High haplotype diversity and low linkage disequilibrium suggest that the polymorphic restriction sites around *Adh* have segregated almost randomly during the history of this *D. pseudoobscura* population and that the effects of inbreeding and periodic reduction of population size have been negligible. The restriction-endonuclease analysis for the *Adh* region of *D. pseudoobscura* stands in sharp contrast to the strong linkage disequilibrium, high levels of insertion/deletion polymorphism, and lower estimates of nucleotide polymorphism found for this same region in *D. melanogaster*. A phylogeny for the *Adh* haplotypes is consistent with an early divergence of *D. miranda*, while *D. persimilis* falls within the cluster of *D. pseudoobscura* haplotypes.

**Introduction**

Analysis of restriction sites allows large regions of DNA to be sampled for nucleotide sequence diversity. The potential of restriction-endonuclease analysis for measuring genetic variability in natural populations has begun to be realized by population geneticists (Avise et al. 1979; Jeffreys 1979; Brown 1980; Wyman and White 1980; Langley et al. 1982). The alcohol-dehydrogenase-gene (*Adh*) locus of *Drosophila* is a model system for studying the processes of molecular evolution. The molecular genetics of *Adh* is well characterized in *D. melanogaster* (Goldberg 1980; Benyajati et al. 1981, 1983; Goldberg et al. 1983; Kreitman 1983). The *Adh* locus in this species has two common protein variants in populations (Johnson and Denniston 1964). Langley et al. (1982) and Aquadro et al. (1986) have carried out detailed restriction analyses of the *Adh* region of *D. melanogaster* and found that any two randomly chosen chromosomes differ at one in every 166 nucleotides and that one in 37 nu-

1. Key words: population genetics, molecular evolution, comparative genetics, polymorphism, insertion/deletion.
2. Current address: Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138.
Address for correspondence and reprints: Dr. Stephen W. Schaeffer, Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138.

© 1987 by The University of Chicago. All rights reserved.
0737-4038/87/0403-0005$02.00

254
cleotides is polymorphic when 48 lines are studied. They also have found strong non-
random associations between the two major Adh electromorphs and restriction sites
that flank the gene. In addition, they have noted that 80% of the second chromosomes
carry one or more sequence length variants in the Adh region.

The comparative genetics of Drosophila species has given considerable insight
into the latters' evolutionary biology. Sturtevant and Tan (1937) used linkage maps
for visible mutants to establish homologies between chromosomes in D. melanogaster
and D. pseudoobscura. McDonald and Avis (1976) noted differences in the devel-
opmental expression of Adh among several Drosophila species groups. Comparison
of DNA sequences and restriction maps of Adh within D. melanogaster and among
closely related species (Langley et al. 1982; Kreitman 1983; Bodmer and Ashburner
1984; Cohn et al. 1984) have clarified evolutionary relationships and revealed evidence
of purifying selection. Comparison of the D. melanogaster sequence to the homologous
sequences of the more distantly related D. pseudoobscura (S. W. Schaeffer and C. F.
Aquadro, unpublished results) have supported these results and highlighted regions
of selective constraint, including a new gene 3' to Adh. We present here a restriction-
endonuclease analysis of the Adh region in D. pseudoobscura, D. persimilis, and D.
miranda to provide estimates of nucleotide variability and divergence within and
among these species and to allow comparison with similar measurements in D. me-
lanogaster and its close relatives.

Material and Methods

Samples

Twenty-two isofemale lines from the pseudoobscura subgroup of the obscura
species group were studied (table 1). Nineteen lines of Drosophila pseudoobscura were
sampled from a single population at the Rainbow Orchard in Apple Hill (AH), Cal-
ifornia, in October 1982. An additional line of D. pseudoobscura from Bogotá (BogER)
was collected in 1979 by Dr. Hugo Hoenigsberg. The D. miranda line was collected
in California by Drs. John and Betty Moore in 1977. The D. persimilis line is a mutant
strain for orange eyes from the Drosophila species stock center. Each isofemale strain
was inbred by brother-sister matings for 20 generations in an effort to make the
Adh region homozygous. Nonetheless, several strains remained heterozygous, since we found
in them a combination of restriction-map patterns from two divergent haplotypes. If
the frequencies of the two haplotypes within a line were extremely different, the two
haplotypes could be resolved, as in the case of AH172 and AH91. However, for two
other lines, the two haplotypes were present in similar frequency and the pattern could
not be resolved into two haplotypes. Hence, these two lines were excluded from the
analysis.

Clone Isolation

Several loci in D. pseudoobscura show alcohol dehydrogenase (ADH) activity on
electrophoretic gels; according to Chambers et al. (1978), the gene homologous to Adh
in D. melanogaster is the Adh-I locus described by Singh (1976). The details of the
cloning and characterization of the Adh locus in D. pseudoobscura have been detailed
by us (unpublished results). From a genomic library we isolated an EMBL4 phage
clon containing a 15.2-kb insert with DNA homologous to the Adh locus of D. me-
lanogaster. This D. pseudoobscura clone was designated Adh6. Langley et al. (1982)
designated the BamHI site in the Adh coding sequence of D. melanogaster as the zero
Table 1

**Adh-Region Haplotypes for 20 Strains of Drosophila pseudoobscura**

| STRAINS | a | b | c | d | e | f | g | h | i | j | k | l | m | n | o | p | q | r | s | t | u | v | w | x | y | z | a' |
| AH38    | - | - | - | - | - | - | - | - | - | + | + | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH130   | - | + | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH135   | + | + | + | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH172a  | - | + | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH172b  | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH154   | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH100   | + | + | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH41    | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH143   | - | + | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH133   | - | + | + | + | - | - | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| AH162   | + | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH165   | + | + | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH49    | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH91a   | - | + | + | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH91b   | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH122   | + | + | + | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH144   | - | + | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH4     | - | + | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AH54    | - | - | - | - | - | - | - | - | - | - | - | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| BogER   | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |

* Restriction site coordinates (see fig. 1) are as follows: a, EcoRI -17.6; b, XbaI -14.3; c, Xho -13.3; d, HindIII -11.3; e, XbaI -9.3; f, XbaI -8.0; g, EcoRI -7.3; h, HindIII -6.4; i, EcoRI -5.6; j, HindIII -5.3; k, HindIII -5.1; l, EcoRI -4.3; m, XbaI -2.4; n, EcoRI -2.4; o, HindIII -1.8; p, HindIII -1.5; q, EcoRI -0.4; r, HindIII 2.1; s, XhoI 3.3; t, EcoRI 3.5; u, HindIII 3.5; v, XbaI 5.8; w, EcoRI 7.0; x, SalI 9.3; y, XhoI 9.5; z, BamHI 13.7; and a', EcoRI 14.1.

position. The homologous BamHI site in *D. pseudoobscura* has been lost (as determined on the basis of sequence data [S. W. Schaeffer and C. F. Aquadro, unpublished results]). Nonetheless, we used the homologous location of the BamHI site in *D. pseudoobscura* as the zero position and numbered bases in the Adh region accordingly. The Adh6 probe hybridizes to sequences extending 16.1 kb upstream from Adh and 2.8 kb downstream from Adh. To examine the region 3' to Adh, we isolated a second clone, Adh8, which consists of the 10.2-kb SalI fragment from the BogER strain in an EMBL3 phage vector (fig. 1). This clone was isolated from an EMBL3 library (Frischauf et al. 1983) constructed from a complete SalI digest of genomic DNA from the BogER strain. The 1.8-kb BamHI fragment from the 3' end of Adh6 was used as a probe to isolate this clone. The protocols used for phage propagation and isolation of DNA from phage are those of Maniatis et al. (1982).

**Restriction Analysis**

We isolated genomic DNA from 1–2 g of flies from each isofemale line according to the protocol of Bingham et al. (1981). Each line was mapped for the following restriction endonucleases, each recognizing a 6-bp sequence: BamHI, EcoRI, HindIII, SalI, XbaI, and XhoI (commercially available from Bethesda Research Laboratories). Total genomic DNA (1–2 μg) was digested with each of the six restriction endonucleases
according to the method described by Maniatis et al. (1982). Restriction fragments were separated by size on 0.8% agarose gels (McDonell et al. 1977). DNA fragments were transferred to nylon filters (Zetabind, commercially available from AMF Cuno) by means of the technique of Southern (1975). Hybridization, washing, and autoradiography of the filters were carried out according to the procedure of Wahl et al. (1979). The Adh region of D. pseudoobscura was probed with the 15.2-kb clone, Adh6, and the 10.2-kb clone, Adh8. These clones probe a 32-kb region extending 16.1 kb and 14.3 kb to the 5' and 3' ends, respectively, of the adult transcript of Adh.

Results
Nucleotide-Substitution Variation

Restriction-endonuclease-site maps for the 32-kb Adh region of Drosophila pseudoobscura, D. persimilis, and D. miranda are shown in figure 1. The 27 variable and 16 monomorphic sites are indicated for the D. pseudoobscura lines. Presence and absence of sites polymorphic within D. pseudoobscura are indicated for each line in table 1. We will refer to these multisite genotypes as haplotypes. For the AH population of D. pseudoobscura, we estimate per-nucleotide heterozygosity (H) and polymorphism (p) to be 0.026 and 0.076, with SDs of 0.005 and 0.015, respectively. These estimates were calculated using equations 7, 11, and 19 from Hudson (1982) and equation 11 from Ewens et al. (1981). Estimates of these parameters by other approaches (e.g., those of Engels [1981] and Nei and Tajima [1981]) give essentially the same results. The heterozygosity estimate for D. pseudoobscura is fourfold higher than that reported for a 13-kb segment of the Adh region in the Rhode Island population of D. melanogaster by Aquadro et al. (1986), who found that H = 0.006 with an SD of 0.001. Other D. melanogaster populations showed levels of variability consistent with that of this Rhode Island sample (Aquadro et al. 1986; Kreitman and Aquadro 1986).

Including only those restriction sites that are within the same 13-kb region sampled

![Restriction maps of the Adh region of Drosophila pseudoobscura, D. persimilis, and D. miranda.](image-url)
by Aquadro et al. (1986) in D. melanogaster gives estimates for D. pseudoobscura that are virtually the same as those for the 32-kb region (H = 0.023 and p = 0.068, with SDs of 0.006 and 0.017, respectively). The differences in variability estimates between D. pseudoobscura and D. melanogaster do not appear to be biased by the restriction enzymes used, since recalculation of enzymes common to both studies yields differences of the same magnitude. Thus, the population of D. pseudoobscura sampled seems to harbor more nucleotide sequence variation in the Adh region than does D. melanogaster.

Sequence length variation was restricted to only two lines, both with insertions ∼6 kb upstream from Adh. The AH69 line had an insertion of ∼50 bp between −7.1 and −7.5 on the restriction map (fig. 1). The AH165 line had an insertion of 200 bp between coordinates −8.1 and −10.6. (Note that these variants are referred to as insertions simply on the basis of comparison to the most common restriction map for these species.) The low level of detectable sequence length variation in D. pseudoobscura is in sharp contrast to the frequency of lines of D. melanogaster that carried an insertion or deletion in the Adh region (80%) (Aquadro et al. 1986). Examination of other regions in D. melanogaster has also revealed substantial levels of insertion/deletion polymorphism (see, e.g., the 87A heat-shock-gene region; Leigh-Brown 1983).

**Haplotype Diversity**

Haplotype diversity measures variability as the expected heterozygosity of multisite genotypes (haplotypes; Nei and Tajima 1981). If only one genotype is present in the population, then the haplotype diversity is zero. If every chromosome is unique, then the haplotype diversity is one. The AH population of D. pseudoobscura has a haplotype diversity of one for the 32-kb Adh region. If we consider only the 13.3-kb region of the D. pseudoobscura restriction map homologous to that examined in D. melanogaster (map coordinates −7.5 to 5.8; Langley et al. 1982; Aquadro et al. 1986), the diversity index decreases only slightly, to a value of 0.985. In contrast, haplotype diversity for restriction-site polymorphism is 0.856 in D. melanogaster. Haplotype diversity for sequence length variation was 0.204 in D. pseudoobscura, again in marked contrast to the situation in D. melanogaster, in which it was 0.917.

**Random Association of Restriction Sites**

Extensive recombination among divergent genotypes and the resultant randomization of variation at adjacent polymorphic sites along the chromosome can contribute to high haplotype diversity. We therefore tested whether restriction sites were associated randomly in haplotypes. Our measure of nonrandom association is D', the ratio of the observed linkage disequilibrium D, to its theoretical maximum, Dmax (Lewontin 1964). All pairwise comparisons between restriction sites were made provided the frequency of the restriction site (+) was >20% and <80%. Restriction sites whose frequencies were >80% or <20% were excluded because the probability of sampling all four gametic types between any two such restriction sites is small, regardless of the extent of random association. Observed values of D are related to the \( \chi^2 \)-statistic with 1 df. Only one of the 28 pairwise comparisons indicated significant nonrandom association (table 2). The null hypothesis of random association may not have been rejected in the bulk of the comparisons because the power of the test was reduced owing to small D values and small sample size (Chakraborty 1984). A comparison with data for D. melanogaster is helpful in this regard. Seven of 45 comparisons had
Table 2

<table>
<thead>
<tr>
<th>Column 1</th>
<th>Column 2</th>
<th>Column 3</th>
<th>Column 4</th>
<th>Column 5</th>
<th>Column 6</th>
<th>Column 7</th>
<th>Column 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>XbaI</td>
<td>XbaI</td>
<td>EcoRI</td>
<td>HindIII</td>
<td>EcoRI</td>
<td>HindIII</td>
<td>EcoRI</td>
<td>EcoRI</td>
</tr>
<tr>
<td>-14.3</td>
<td>-8.0</td>
<td>-2.4</td>
<td>-1.5</td>
<td>3.5</td>
<td>3.5</td>
<td>7.0</td>
<td>14.1</td>
</tr>
</tbody>
</table>

1... 0.208 0.050 -0.525 -0.472 0.136 0.367 0.050 0.028
2... 0.008 0.028 0.379 0.186 0.260 -0.321 -1.000 0.050
3... -0.058 0.069 -0.006 -0.156 0.208 -0.050 -0.525 0.050
4... -0.047 0.036 0.058 -0.039 -0.472 -1.000 0.046 -0.406
5... 0.017 0.055 -0.089 0.042 -0.094 -0.406 -0.406 0.050
6... 0.061 -0.025 0.008 -0.006 -0.100 -0.036 0.050 0.008
7... 0.008 -0.078 0.008 -0.058 -0.047 -0.036 0.008 0.008

NOTE.—D' values are above the diagonal; D values are below the diagonal. A value of 0 indicates completely random association; a value of -1 or 1 represents complete association.

*χ²-Value associated with this value of D' significant at P < 0.05.

nonrandom associations for a data set in D. melanogaster that was of comparable size (Langley et al. 1982). A sample of 22 chromosomes from a D. melanogaster population in Rhode Island also had many more significant nonrandom associations than did D. pseudoobscura (Aquadro et al. 1986). Thus, the higher number of nonrandom associations in D. melanogaster were not simply the result of a larger sample size increasing the power of the test.

In the absence of recurrent or reverse mutation in an infinite-sites model, the only way to generate all gametic types (+ +, + -, - +, - -) is by means of recombination (Hudson and Kaplan 1985). We see all four gametic types in 25 of 28 pairwise comparisons of restriction sites in D. pseudoobscura. Thus, our data indicate that recombination has generated all gametic types, provided that the probabilities of parallel losses and parallel gains of restriction sites are small. These results stand in sharp contrast to the strong nonrandom association observed among restriction sites and ADH allozymes in D. melanogaster (Aquadro et al. 1986).

Phylogeny

The recombination-based randomization of sequence variation in the Adh region of D. pseudoobscura makes construction of a maximum-parsimony tree (Fitch 1977) for these haplotypes extremely difficult. Given the small number of polymorphic sites relative to the extensive amount of recombination, distinguishing between parallel/convergent events and recombination events becomes very difficult (Aquadro et al. 1986; Stephens 1986). Therefore, we have used a phenetic approach to summarize the relationships among the 20 lines of D. pseudoobscura, the D. persimilis line, and the D. miranda line (fig. 2). This phenogram was produced by applying the unweighted-pair-group method of analysis (UPGMA) clustering algorithm (Sneath and Sokal 1973, pp. 230-234) to a matrix of pairwise distance estimates calculated (using eq. 8 of Nei and Li 1979) from the restriction-map data. The D. miranda haplotype is the most divergent, differing from the D. pseudoobscura and D. persimilis lines by an estimated 4% of its nucleotides. Extensive divergence also exists among the D. pseudoobscura haplotypes, with divergence among many haplotypes being ≥2%-3% (fig. 2). Perhaps
surprisingly, both the *pseudoobscura* line from Bogotá and the *D. persimilis* line cluster within the relatively diverse *pseudoobscura* haplotypes. The magnitude of differentiation among these three *obscura*-group species closely parallels the divergence in the *Adh* region among *D. melanogaster*, *D. simulans* and *D. mauritiana* of the *D. melanogaster* species group (Langley et al. 1982).

**Discussion**

Allozyme studies have indicated roughly equivalent levels of protein polymorphism for populations of *Drosophila pseudoobscura* and *D. melanogaster* (Prakash et al. 1969; Kojima et al. 1970; Singh and Coulthart 1982). These data suggest that the average levels of nucleotide polymorphism might also be similar in these two species. Our data, however, show that *D. pseudoobscura* populations harbor substantially higher levels of restriction-site polymorphism than do *D. melanogaster* populations, at least for the *Adh* region. This single population of *D. pseudoobscura* contains more nucleotide variation than do samples of *D. melanogaster* taken from around the world (Kreitman 1983; Aquadro et al. 1986; Cross and Birley 1986). This direction of the difference is particularly surprising since ADH is reported to be electrophoretically monomorphic in *D. pseudoobscura* while *D. melanogaster* harbors two common and several rare ADH protein variants (see, e.g., Johnson and Denniston 1964; Singh 1976; Gibson et al. 1982).

Despite the high level of restriction-site polymorphism in *D. pseudoobscura*, insertion/deletion variability was rare. This again contrasts strongly with the situation in *D. melanogaster*, in which sequence length variation, much of which is due to the insertion of transposable elements, is extensive (see, e.g., the *Adh* region [Aquadro et
al. 1986] and the 87A heat-shock-gene region [Leigh-Brown 1983]). A possible explanation for the paucity of sequence length polymorphism in \textit{D. pseudoobscura} is that this species has fewer numbers and/or families of transposable elements than does \textit{D. melanogaster}. When the \textit{D. pseudoobscura} genome is probed with transposable elements from \textit{D. melanogaster}, only a few families of sequences are represented (Martin et al. 1983; Brookfield et al. 1984). Of course, this may only indicate rapid divergence of these elements during the time since these two species diverged. It is also possible that \textit{D. pseudoobscura} has a largely different set of elements than that found in \textit{D. melanogaster}. A characterization of the nomadic sequence families in \textit{D. pseudoobscura} is needed to distinguish between these hypotheses.

An explanation for the apparently higher level of nucleotide polymorphism observed in \textit{D. pseudoobscura} is not immediately evident. Sampling across a number of genetically differentiated demes could inflate estimates of genetic variation. However, the expected nonrandom associations among restriction sites were not observed. In fact, the presence of each haplotype only once in the sample suggests that the effects of drift and inbreeding due to population subdivision and bottlenecks were negligible in the AH population of these flies.

The AH population appears to be old because a sufficient period of time has passed to allow recombination to generate all four gametic types for most pairwise comparisons of restriction sites in the \textit{Adh} region. This is evidenced by the extremely high diversity of haplotypes and the lack of linkage disequilibrium. These results differ markedly from those obtained in \textit{D. melanogaster}, in which restriction sites in the 13-kb region around \textit{Adh} are in strong nonrandom association with both each other and the \textit{Adh} allozyme (Aquadro et al. 1986). These patterns in \textit{melanogaster} may reflect the recent increase in frequency of the \textit{AdhF} allele and the hitchhiking of adjacent sequences (Aquadro et al. 1986). Perhaps a more meaningful comparison with \textit{D. pseudoobscura}, therefore, would be made by removing the \textit{AdhF} allele from the \textit{D. melanogaster} data set. Doing this by analyzing only \textit{AdhS} in \textit{D. melanogaster}, we still find numerous significant nonrandom associations over the 13-kb region. Thus, the two species clearly differ in level and organization of DNA sequence variation over the \textit{Adh} region.

Under the infinite-site model of neutral mutations, restriction-map variation can be used to estimate the population parameter $4N_e\mu$, which is equal to per-nucleotide heterozygosity (see, e.g. Ewens et al. 1981; Nei and Tajima 1981; Hudson 1982); $N_e$ is the effective population size, and $\mu$ is the substitution rate per nucleotide site. We have shown that the nucleotide heterozygosity for \textit{D. pseudoobscura} is approximately four times that for \textit{D. melanogaster}. Taken at face value, this difference suggests a higher rate of restriction-site change and/or a larger effective population size for \textit{D. pseudoobscura} than for \textit{D. melanogaster}. Clearly, more extensive sampling of populations—using larger numbers of restriction sites and, preferably, direct sequencing—is needed before this conclusion can be accepted.

\textit{Drosophila persimilis}, \textit{D. miranda}, and \textit{D. pseudoobscura} are sibling species, while the population in Bogotá has been described as a subspecies of \textit{D. pseudoobscura} (Ayala and Dobzhansky 1974). Consistent with studies of morphology, chromosomal arrangements, protein polymorphism, and mating preference (reviewed in Lakovaara and Saura 1982), \textit{D. miranda} is the most divergent \textit{Adh} haplotype (fig. 2). However, divergence among \textit{D. pseudoobscura} haplotypes is also large, with one line (AH162) nearly as divergent from other \textit{pseudoobscura} lines as is the \textit{miranda} haplotype. There are no clear clusters of alleles within \textit{pseudoobscura} but rather a continuum of levels
of divergence. While *D. persimilis* is clearly a distinct species on the basis of reproductive isolation (Powell 1983, and citations therein), the one haplotype from this species that we analyzed is less divergent from many *pseudoobscura* alleles than are many of the other *pseudoobscura* alleles. This pattern may have resulted from introgression of *pseudoobscura* alleles into *persimilis*, as has been proposed for mitochondrial genes (Powell 1983). However, we feel that the retention of ancestral polymorphism is a simpler and more reasonable explanation of our data.

Virtually all proposed mechanisms of cladogenic speciation involve the reproductive isolation, from other members of an ancestral taxa, of a subsample of individuals of a species. It is possible that the subsample may, in fact, often be quite large. Nonetheless, the allelic composition of the new species will contain a subset of those alleles segregating in the ancestral species. In fact, much of the initial divergence between closely related taxa may be due to alternative fixation or frequency shifts of alleles at ancestrally polymorphic loci. If substantial polymorphism exists in the ancestral gene pool, then some of the alleles captured in the new species may have separate evolutionary histories that substantially predate the cladogenic event (Tajima 1983; Takahata and Nei 1985). In a highly polymorphic species such as *D. pseudoobscura*, population-size bottlenecks and other mechanisms that reduce sequence variability in a species do not appear to have had a substantial effect. Hence, many of the polymorphisms probably predate the *persimilis-pseudoobscura* speciation. Thus, for closely related taxa that are highly polymorphic and/or have evolved from a highly polymorphic ancestor, allele phylogenies for single loci may not accurately reflect species phylogenies, since allele divergences may substantially predate the speciation event. Similar results have been discussed for mitochondrial DNA in the rodent *Peromyscus* (Avise et al. 1983; see also Neigel and Avise 1986).

In summary, our data on levels and pattern of restriction-site variation for the *Adh* region indicate that the AH population of *D. pseudoobscura* has persisted as a large, random mating population for a very long time. The retention of a large amount of ancestral polymorphism, as reflected in our phylogeny of *Adh*-region haplotypes from *D. pseudoobscura*, *D. persimilis*, and *D. miranda*, also lends support to the hypothesis of the long-term stability of these populations. This conclusion is consistent with the distribution of electromorph frequencies at the esterase-5 and xanthine dehydrogenase loci in other California populations of this species (Keith 1983; Keith et al. 1985). Our finding of substantially different levels and organization of DNA variation in *pseudoobscura* compared with those observed for the *Adh* and other gene regions in *melanogaster* was unexpected. Whether the lower levels of variability and higher nonrandom associations observed for *D. melanogaster* represent a nonequilibrium situation for this species or whether they reflect differences between *melanogaster* and *pseudoobscura* in population size, population structure, or the action of selection will be interesting to investigate. Of particular interest will be comparison of restriction-map variation in other gene regions—and DNA sequencing of several alleles at the *Adh* locus itself.

**Acknowledgments**

We would like to thank Dick Hudson, Clay Stephens, and an anonymous reviewer for helpful suggestions on this manuscript. We also thank Chuck Langley for his encouragement and stimulating discussions.
LITERATURE CITED


Masatoshi Nei, reviewing editor

Received June 6, 1986; revision received October 31, 1986.
Phosphoglucose isomerase (PGI) allozymes were isolated from the wing-polymorphic water strider, *Limnoporus canaliculatus*, and were characterized biochemically with respect to temperature-dependent kinetic and thermostability properties. At higher temperatures, the allozymes exhibited significant differences in Michaelis constant ($K_m$) values for substrates of both the forward and reverse reaction directions. Results were consistent with expectations of adaptive kinetic differentiation based on the latitudinal variation of PGI allele frequencies. PGI genotypes also differed with regard to maximal velocity ($V_{max}$)/$K_m$ ratios at higher temperatures. These differences were due primarily, if not exclusively, to allozyme-dependent variation in $K_m$ values. The allozymes also exhibited dramatic differences in thermostability. However, no thermostability differences were observed when the substrate analogue 6-phosphogluconate was present in the incubation medium. The data from this study, together with data from *Mytilus edulis* and *Metridium senile* on temperature-dependent kinetic variation among PGI allozymes, form a consistent picture of natural selection influencing the clinal variation of alleles at this locus in these three phylogenetically distant organisms. More definitive support of this hypothesis, however, must await additional studies on the physiological effects of the allozymic variation as well as direct measurements of fitness differences among the enzyme genotypes.

**Introduction**

A major goal of population genetics during the past 20 years has been to identify the forces acting on allozyme variation (Lewontin 1974; Nei 1975; Ayala 1976; Nei and Koehn 1983). Research on this topic has essentially attempted to determine the relative importance of selection versus mutation-drift in controlling the frequencies of allozyme variants in natural populations. An adequate demonstration that a specific enzyme polymorphism is influenced by selection is a difficult task requiring information on the biochemical properties of allozymes, the physiological consequences of allozymic variation, and the mechanisms whereby the physiological variation is translated into fitness differences (Clarke 1975; Koehn 1978). Biochemical studies of the kinetic and physical properties of allozymes are a key component in these multilevel investigations (Koehn et al. 1983; Zera et al. 1985). Biologically meaningful differences must be
demonstrated among allozymes before selection can justifiably be invoked as influencing an enzyme polymorphism in natural populations.

Both the reliability and the biological relevance of any observed biochemical variations among allozymes are contingent on a number of factors, including the kinetic parameters chosen for study, the conditions under which these parameters are estimated, and the interpretations drawn from the primary enzymological data. Although a large number of biochemical characterizations of allozymes have been reported during the past 15 years, many of these studies have shortcomings in one or more of the aspects mentioned above (Koehn et al. 1983; Zera et al. 1985). Indeed, there is a paucity of reliable, biologically meaningful data on the physical and kinetic properties of allozymes. Obtaining such data remains one of the most important tasks of current allozyme research (Zera et al. 1985).

For a variety of reasons, phosphoglucose isomerase (PGI) polymorphisms are useful experimental systems for addressing questions of allozymic adaptation. First, there is considerable background information on the structure, kinetic properties, and metabolic roles of PGIs from a variety of organisms (for review, see Noltman 1972). Second, the enzyme is highly polymorphic in natural populations of many plant and animal species, and in many cases gene-frequency clines associated with latitude have been reported (Williams et al. 1973; Koehn et al. 1976; Corbin 1977; Powers and Place 1978; Hoffman 1981b). These data suggest that PGI polymorphisms appear to be influenced by natural selection and also allow specific a priori predictions to be made regarding adaptive differences among allozymes in temperature-dependent kinetic and physical properties. Last, in recent years, both biochemical and population-genetic data have been reported for PGI polymorphisms in several species (Watt 1977, 1983; Hoffman 1981a, 1981b, 1983; Hall 1983, 1985b). Thus, additional studies of PGI polymorphisms afford the opportunity to investigate the comparative biochemical population genetics of an enzyme locus to a degree not possible with allozymes at most other loci.

In an earlier study, PGI alleles in the wing-polymorphic water strider, *Limnoporus canaliculatus*, were found to exhibit a steep latitudinal cline (Zera 1984; also see below), thus making this a convenient experimental system for studying the thermal adaptation of PGI allozymes. In addition, Pgi allele frequencies differed substantially between winged and wingless morphs throughout the entire range of *L. canaliculatus* (Zera 1984). This suggested that PGI allozymes might be differentially selected in the two morphs (analogous to differential selection between the sexes), possibly as a consequence of kinetic differences related to flight biochemistry. Thus, the PGI polymorphism in *L. canaliculatus* was also viewed as a useful experimental system for investigating the functional adaptation of allozymes to different metabolic contexts.

The goal of the present study was to estimate various steady-state kinetic and thermostability parameters of PGI allozymes under physiologically realistic conditions and under a range of biologically relevant temperatures. Correlation between temperature-dependent variation in enzymatic characteristics among the allozymes and latitudinal allele frequency variation would be the first step in implicating selection as a factor affecting the PGI polymorphism in *L. canaliculatus*. Comparison of these data to biochemical data on other clinally varying PGI allozymes would potentially identify general patterns of allozymic adaptation at this locus. Results of a companion study of the differential inhibition of PGI allozymes by pentose-shunt metabolites and its role in the association between wing polymorphism and PGI polymorphism will be reported elsewhere (A. J. Zera, accepted).
Material and Methods
Background on the PGI Polymorphism

The PGI polymorphism in *Limnoporus canaliculatus* consists of two common and several rare (frequency < 0.01) allozymes. The two common PGI allozymes are encoded by codominant alleles at a single locus in linkage group 1 (Zera 1984, and accepted). PGI allele frequencies exhibit an interesting pattern of latitudinal variation (figs. 1, 2A). The $Pgi^F$ allele increases in frequency from a low of ~0.3 in central Maine, the northernmost population sampled, to a high of ~0.8 in southern Georgia–northern Florida. South of this point, the cline reverses and the frequency of $Pgi^F$ declines linearly to 0.3 in southern Florida. The reversal of the $Pgi$ cline occurs at $30^\circ$ north latitude, precisely at the north-Florida suture zone (Remington 1968). This is

**Fig. 1.—** Location of populations of *Limnoporus canaliculatus* sampled in the present study.
Kinetics of PGI Allozymes from *L. canaliculatus* 269

North Florida Suture Zone (Remington, 1968)

![Graph showing latitudinal variation of PGI allele frequencies](image)

**Fig. 2.**—Latitudinal variation of (A) *Pgi*<sup>F</sup> allele frequencies, (B) *Idh*<sup>B</sup> allele frequencies, and (C) *Pgi* allele-X wing-morph associations in fall samples of *Limnoporus canaliculatus*. Data points represent population samples (*N* = 50–400 individuals per sample) taken in 1979 or 1981 (●) or pooled samples from both years (○). □ = The weighted mean *Pgi*<sup>F</sup> allele frequency for samples in which *Pgi*<sup>F</sup> differed significantly between long-winged and wingless morphs. * = A significant difference (*P* < 0.05 and usually <0.005) in *Pgi*<sup>F</sup> allele frequency between long-winged and wingless morphs within individual population samples as determined by *χ²* contingency analyses. Lines represent results of regression analyses of untransformed allele frequency on latitude and are significant (*P* < 0.001) in each case, except for *Idh*<sup>B</sup> above 30° north latitude. Regressions of arcsine-transformed allele frequencies were also significant (*P* < 0.001) in each case. Because of the sharp discontinuity of allele frequencies at 30° north latitude, regression analyses were performed separately for samples north and south of this point. See text for additional details.

A narrow belt of hybridization of many plant and animal species, one presumably due to the contact of previously isolated peninsular Floridian populations or species and their continental near relatives. Besides the sharp reversal of the *Pgi* allele-frequency cline, two other characters in *L. canaliculatus* also change abruptly in this area: latitudinal variation of allele frequencies at the unlinked (Zera 1984) *Idh* locus (fig. 2B) and *Pgi* allele-X wing-morph associations (fig. 2C). The abrupt change in each of these
three characters indicates that populations of *L. canaliculatus* north of 30° north latitude are strongly differentiated from populations south of this point, at least at some loci.

A sufficient study of the biochemical mechanisms underlying the entire pattern of latitudinal variation of the *Pg* locus in *L. canaliculatus* requires isolation and characterization of PGI allozymes from both Floridian and continental populations. This is necessary because one or more of the PGI allozymes in Floridian populations, while electrophoretically indistinguishable from its counterpart in continental populations, could nevertheless be functionally differentiated (i.e., a cryptic variant), resulting in the reversal of the PGI allele-frequency cline. The present study is more limited in scope and focuses exclusively on the portion of the PGI cline north of the north-Florida suture zone, where the frequency of *Pgi* is negatively correlated with latitude. Thus, in the present study, PGI-FF is considered the hypothetically warm-adapted enzyme while PGI-SS is considered the hypothetically cold-adapted enzyme.

**Chemicals**

Proteins, enzymes, substrates, coenzymes, and buffers used in this study were obtained from Sigma Chemical Company, with the exception of fructose-6-phosphate (F-6-P), which was purchased from Calbiochem-Behring (lot 1183258). F-6-P from Sigma contained an unknown substance that strongly and differentially inhibited PGI allozymes from *L. canaliculatus* in a manner similar to that recently reported for PGI allozymes from *E. coli* (Dykhuizen and Hartl 1983). All other organic and inorganic chemicals were reagent grade and were obtained from Fisher Chemical Company.

**Animals**

*Limnoporus canaliculatus* collected from the Nissequogue River State Park, Long Island, New York, were used as the source of PGI allozymes. Stocks of the three *Pg* genotypes—*Pgi* FF, *Pgi* FS, and *Pgi* SS—were constructed, and individuals of known *Pg* genotype were stored frozen at -60°C until homogenization for enzyme purification. Laboratory rearing conditions were as previously described (Zera 1981).

**Enzyme Preparation**

PGI allozyme preparations used in this study were purified by a combination of ion exchange, gel sieving, and hydrophobic interaction chromatography, the details of which are reported elsewhere (Zera 1984, and accepted). Although the preparations were highly purified (>800 fold), and PGI was the major protein present, the enzymes could not be purified to homogeneity. This was mainly due to the limited amount of tissue available for enzyme extraction and to the instability of the enzyme during the final ultrafiltration step.

**Enzymc Assay**

Initial reaction velocities were measured by following the change in absorbance at 340 nm using a Gilford model 2400-2 UV/VIS recording spectrophotometer equipped with a ThermoSet temperature-control device (±0.4 degrees C). The spectrophotometer was interfaced with a Motorola M6800 microcomputer that recorded the optical density readings and computed initial rates.

All assays were done using Na-MOPS buffer at 0.1 ionic strength (I). Initial rates were determined in the gluconeogenic direction (F-6-P → G-6-P) using an assay mixture containing Na-MOPS buffer at various pHs, 8 mM MgCl2, 1 mM NADP+, various
F-6-P concentrations, and 0.5–2.0 units of glucose-6-phosphate dehydrogenase (G-6-PDH), depending on the temperature, in a total volume of 0.5 ml. Initial rates were determined in the glycolytic direction (G-6-P $\rightarrow$ F-6-P) using a modified version of the multiply coupled enzyme assay of Tilly et al. (1974). The assay mixture consisted of Na-MOPS buffer, 8 mM MgCl$_2$, 1 mM NH$_4$Cl, 0.15 mM NADH, 0.2 mM ATP, 0.1% β-mercaptoethanol, various G-6-P concentrations, and enough of the coupling enzymes (phosphofructokinase [PFK], α-glycerophosphate dehydrogenase [α-GPDH], triosephosphate isomerase [TPI], and aldolase [ALD]) to ensure that the coupling enzymes were not rate limiting (e.g., 1 unit of PFK, 5 units of TPI, and 0.5 units each of α-GPDH and ALD when rates were measured at 20 C). The standard assay used in the routine measurement of PGI activity (e.g., during thermostability studies, specific activity determinations, etc.) consisted of 0.1 l Na-MOPS, pH 7.1, at 20 C, 1 mM F-6-P, 1 unit of G-6-P, 8 mM MgCl$_2$, and 1 mM NADP$^+$. Assays in both directions were initiated by the addition of 10 µl or 20 µl of suitably diluted PGI stock solution to the assay cocktail, which had been preincubated at the assay temperature for a few minutes. In all cases <5% of the substrate was consumed during initial rate measurements in order to operate within the restrictions of steady-state kinetics (Fromm 1975). Initial G-6-P $\rightarrow$ F-6-P and F-6-P $\rightarrow$ G-6-P rates near maximal velocity ($V_{\text{max}}$) were estimated by unweighted linear least-squares analysis of velocity measurements. Since the Michaelis constant ($K_m$) for F-6-P was low (<40 µM), it was not possible to both accurately estimate initial velocities using a linear least-squares method and adhere to the restrictions of steady-state kinetics (consumption of <5% of the substrate) when the concentration of F-6-P was near or lower than the $K_m$. Consequently, the progress-curve analysis of Waley (1981; also see Hall and Koehn 1983) was used to measure initial F-6-P $\rightarrow$ G-6-P rates under subsaturating conditions. Stock solutions of substrates and cofactors were prepared according to the method of Lowry and Passonneau (1972) and were stored frozen at $-70$ C until use. Prior to kinetic studies all substrate concentrations were checked spectrophotometrically.

Kinetic Parameters Estimated

$K_m$ and $V_{\text{max}}$ values were estimated for the PGI allozymes at four temperatures—10, 20, 30, and 35 C—that bracketed the temperature range likely to be encountered by *L. canaliculatus* in natural populations. *Limnoporus canaliculatus* has a wide seasonal range and can be found continuously on ponds in New England from April until October (Zera 1980). In the spring and fall *L. canaliculatus* are commonly observed on the water surface at air temperatures >10 C—but uncommonly below this temperature (A. J. Zera, unpublished observations). Thirty-five degrees centigrade was considered to be within the upper limit in temperature experienced by populations in midsummer, especially in southern populations. The $V_{\text{max}}$ and $K_m$ data were subsequently used to estimate $V_{\text{max}}/K_m$ ratios (see below).

Temperature-dependent kinetic studies were done using a variable pH protocol; that is, pH of the assay buffer was adjusted to vary at different temperatures in accordance with the $pK_a$ of imidazole. This procedure mimics the normal intracellular variation of pH with temperature and thus represents a more physiologically realistic experimental protocol than assaying enzyme activity at a constant pH at different temperatures (Somero 1978, 1981).
Experimental Design of Kinetic Studies and Data Analysis

Because the procedures used in the collection and analysis of kinetic data in this study are not widely known, they will be briefly outlined (see Hall and Koehn 1983 for a more detailed discussion and references). The experimental design used here was based on the design recently advocated by Duggleby (1979) and Endrenyi and Chan (1981). This method consists of measuring a large number of rates (7–12 in this study) at each of several substrate concentrations. The number of substrate concentrations at which the rates are measured is equal to the number of kinetic parameters to be estimated (e.g., estimates of $K_m$ and $V_{max}$ require rates to be measured at two substrate concentrations). The optimal experimental design for $K_m$ and $V_{max}$ estimates consists of measuring rates at (1) a high substrate concentration, as close to $V_{max}$ as possible without encountering substrate inhibition, and (2) a low substrate concentration, as close to the $K_m$ as possible.

Although this method results in a more accurate estimate of the kinetic parameters than the traditional method of measuring a smaller number of rates at each of five to seven substrate concentrations, it cannot be used to validate or reject a particular kinetic model (e.g., to determine whether an enzyme obeys Michaelis-Menten hyperbolic saturation kinetics). Since the statistical analysis of data obtained using this method required prior knowledge of the kinetic model of the reaction, preliminary kinetic studies were done to verify that PGI from *L. canaliculatus* exhibits hyperbolic saturation kinetics. It should be noted that PGI isolated from a variety of sources invariably exhibits simple Michaelis-Menten kinetics (see references in Noltman 1972). In preliminary studies of PGI from *L. canaliculatus*, initial rates were measured in the traditional manner in both directions at seven different substrate concentrations, ranging from 0.5 to ~10 times the $K_m$. Double reciprocal plots were constructed, and the lines were examined visually for deviations from linearity.

Enzyme kinetic data were analyzed using the biweight regression technique of Cornish-Bowden and Endrenyi (1981). This is a robust regression technique that requires minimal assumptions about the error structure of the data and mitigates the effects of outliers. The data were analyzed by a computer program written by A. Cornish-Bowden (personal communication) that estimates the various kinetic parameters, their SEs, and df's. As a consequence of the weighting procedure utilized in this method, sample sizes and their resulting df's are nonintegers (e.g., see table 1).

Statistical testing of kinetic parameters was done using analysis of variance (ANOVA) or (for samples containing heterogeneous variances) approximate tests of the equality of means with heterogeneous variances (Sokal and Rohlf 1981). These tests were done after reconstituting the variances from the output data. ANOVA of the kinetic parameters for the three genotypes at a particular temperature was followed by an a priori comparison of the means of the two homozygotes.

Protein Determination

Protein concentrations were determined using the Bio-Rad microassay procedure, which is based on the Coomassie blue G-250 dye-binding assay of Bradford (1976). Bovine serum albumin (BSA) was used as the standard.

Specific Activities

Specific activities were determined for each of the PGI genotypes by means of the standard enzyme assay. Specific activities were measured in each sex and morph
Kinetics of PGI Allozymes from *L. canaliculatus*

Table 1
F-6-P and G-6-P $K_m$ Values (µM) for PGI Allozymes from *Limnoporus canaliculatus*

<table>
<thead>
<tr>
<th>ENZYME AND TEMPERATURE</th>
<th>$K_m$ (µM)</th>
<th>RESULTS OF SIGNIFICANCE TESTS*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SS</td>
<td>FS</td>
</tr>
<tr>
<td>35°C</td>
<td>246 (4), 11.8</td>
<td>274 (19), 11.3</td>
</tr>
<tr>
<td>30°C</td>
<td>212 (10), 14.1</td>
<td>220 (6), 11.9</td>
</tr>
<tr>
<td>20°C</td>
<td>136 (4), 12.3</td>
<td>138 (6), 14.8</td>
</tr>
<tr>
<td>10°C</td>
<td>139 (5), 19.7</td>
<td>ND</td>
</tr>
</tbody>
</table>

F-6-P:

| 35°C                   | 81.2 (4.5), 11.7 | 75.9 (3.1), 15.8 | 61.1 (2.6), 15.9 | All: $P < 0.001$; FF, SS: $P < 0.005$ |
| 30°C                   | 63.2 (2.6), 15.8 | 69.2 (3.3), 19.2 | 53.5 (2.1), 22.7 | All: $P < 0.005$; FF, SS: $P < 0.005$ |
| 20°C                   | 38.3 (1.2), 14.0 | ND | 42.8 (1.5), 22.5 | FF, SS: $P < 0.05$ |
| 10°C                   | 40.3 (1.3), 22.4 | 41.6 (1.8), 14.4 | 40.1 (1.7), 17.0 | All: $P > 0.1$; NS |

NOTE.—NS = Not significant; ND = no data.

*a* See Material and Methods for details of statistical tests of differences among $K_m$ values.

*b* Fractional values are the consequence of the biweighting procedure (see Material and Methods).

**Type** (long winged and wingless) to eliminate potential confounding influences of these variables on variation in activity among *Pgi* genotypes.

**Thermal Denaturation Study**

The thermal denaturation experiment consisted of incubating 200 µl of each allozyme solution at various temperatures for 5 min in a parafilm-sealed test tube. After the incubation period, the tubes were placed in an ice bath and were assayed for PGI activity, in triplicate, using the standard assay. Percent loss in enzyme activity was determined by comparing the PGI activity of solutions incubated at various temperatures with that of a control solution kept on ice. There was no detectable loss in activity of any of the PGI allozymes over several days when diluted in the buffer mentioned above and kept on ice. Prior to these studies, appropriately diluted allozyme solutions were dialyzed against two changes of 200 × vol NaMOPS, 0.1 M, pH 7.1 at 20°C, containing 0.1% β-mercaptoethanol. This was done to remove F-6-P and G-6-P from the allozyme stock solutions and to thus eliminate possible spurious thermostability differences due to differential stabilization of allozymes by the substrates in the incubation solutions. The G-6-P and F-6-P in the allozyme stock solutions were due to the substrate-DEAE ion-exchange chromatographic step, the last step in the purification of the allozymes. All allozyme solutions were dialyzed in the same vessel. Postdialysis enzyme solutions were diluted 1:1 with dialysis buffer containing 1 mg BSA/ml.

The thermostability of PGI allozymes was also monitored in the presence of the substrate analogue and competitive inhibitor, 6-phosphogluconate (6-PG). Assays with inhibitor were done at a concentration of 436 µM 6-PG. The $K_f$ for 6-PG is $\sim 30$ µM at 30°C (Zera 1984, and accepted).

**Computation and Comparison of Relative $V_{max}/K_m$ Ratios**

Differences among the allozymes' $V_{max}$ values observed in this study were the consequence of both biologically relevent factors (e.g., variation in $k_{cat}$ and/or enzyme
concentration) and biologically irrelevant factors (e.g., differential loss of enzyme activity during purification, differences in yields of the various allozyme purifications, etc.). The following method was used to standardize \( V_{\text{max}} \) values among the allozymes to eliminate the biologically irrelevant factors in the variation of \( V_{\text{max}} \) among the allozymes. This would then allow biologically meaningful comparisons of \( V_{\text{max}}/K_m \) ratios to be made among the allozymes. \( V_{\text{max}}/K_m \) ratios were not computed for PGI-FS since, at several temperatures, kinetic constants were not obtained for this enzyme.

Field-collected animals were assayed for genotype-dependent PGI activity by means of the standard assay (at substrate concentrations >20 \( \times \) \( K_m \)) in order to measure velocities essentially at \( V_{\text{max}} \) (velocity > 0.95 \( V_{\text{max}} \) at this substrate concentration). No differences in \( V_{\text{max}} \) were observed among the \( Pgi \) genotypes (table 2). Consequently, the \( V_{\text{max}} \) values for the genotypes obtained in the kinetic studies could be standardized to a common value at the temperature at which the field-collected animals had been acclimated. However, the precise temperature to which the field-collected gerrids had been acclimated was not known. Consequently, two different sets of relative \( V_{\text{max}}/K_m \) values were computed, one in which the \( V_{\text{max}} \) values were standardized at 10 C and another in which the \( V_{\text{max}} \) values were standardized at 20 C. These were considered reasonable values with which to bracket the range in temperatures at which the fall (September 26, 1983)-collected \( L. \) canaliculatus would have been acclimated. Standardization at either value gave virtually identical results. It should be noted that these estimates of \( V_{\text{max}}/K_m \) do not give SEs, thus precluding statistical tests of differences among the allozymes.

Results

Michaelis-Menten Kinetics of PGI

Traditional Lineweaver-Burke plots of the \( 1/v \) versus \( 1/s \) data for both the forward and reverse reactions were constructed (figs. 3, 4) in order to obtain background in-

Table 2

<table>
<thead>
<tr>
<th>MORPH AND SEX</th>
<th>SPECIFIC ACTIVITY (SE) OF PGI GENOTYPE (( \mu \text{mol/min/mg protein} \times 10^4 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FF</td>
</tr>
<tr>
<td>Winged:</td>
<td></td>
</tr>
<tr>
<td>Male ........</td>
<td>2,173 (149)</td>
</tr>
<tr>
<td>N = 6</td>
<td></td>
</tr>
<tr>
<td>Female ......</td>
<td>2,080 (156)</td>
</tr>
<tr>
<td>N = 5</td>
<td></td>
</tr>
<tr>
<td>Wingless:</td>
<td></td>
</tr>
<tr>
<td>Male ........</td>
<td>2,325 (120)</td>
</tr>
<tr>
<td>N = 7</td>
<td></td>
</tr>
<tr>
<td>Female ......</td>
<td>2,060 (0)</td>
</tr>
<tr>
<td>N = 1</td>
<td></td>
</tr>
</tbody>
</table>

**Note.**—Results of the three-way ANOVA of specific activity data: variation among morphs, \( F(1, 101) = 3.99, P < 0.05; \) variation among PGI genotypes, sexes, and all two- and three-way interactions were nonsignificant.
FIG. 3.—Lineweaver-Burke double-reciprocal plot of PGI-SS from *Limnoporus canaliculatus*. The substrate was G-6-P (see Material and Methods); ○ = Two superimposed points.

FIG. 4.—Lineweaver-Burke double-reciprocal plot of PGI-SS from *Limnoporus canaliculatus*. The substrate was F-6-P (see Material and Methods); ○ = Two superimposed points.
formation on the reaction mechanism of PGI from *Limnoporus canaliculatus*. This is a prerequisite for using the kinetic-parameter estimation procedure of Duggleby (1979) and the computer program of Cornish-Bowden. As can be seen, the $1/v$ versus $1/s$ data fit well a straight line for both reaction directions. These data thus provide no evidence that PGI from *L. canaliculatus* differs from other PGIs that invariably exhibit simple hyperbolic saturation kinetics (Noltman 1972). All subsequent kinetic analyses were done under the assumption that PGI from *L. canaliculatus* exhibits simple hyperbolic saturation kinetics.

$K_m$ Values

The $K_m$ values when G-6-P was used as the substrate ranged from a low of $\sim 130 \mu$M at 10 C to a high of 250–275 $\mu$M at 35 C. The $K_m$ values for G-6-P exhibited by the PGI allozymes were virtually identical at 10 and 20 C but increasingly diverged at 30 and 35 C (table 1). The $K_{m(G-6-P)}$ differed significantly between the PGI-FF and PGI-SS enzymes at 30 and 35 but not at 10–20 C. At all temperatures, the PGI-FF enzyme exhibited a lower $K_m$ than did the PGI-SS. Interestingly, the $K_{m(G-6-P)}$ for the PGI-FS enzyme was not intermediate between the $K_{m(G-6-P)}$ values of PGI-FF and PGI-SS but was slightly higher than that of PGI-SS at all temperatures.

The $K_m$ values for F-6-P ranged from a low of 40 $\mu$M at 10 C to a high of 80 $\mu$M at 35 C. Similar allozymic differences were observed for the F-6-P $K_m$ values as were observed for the G-6-P $K_m$ values (table 1). Except at 20 C, the PGI-FF enzyme exhibited a lower $K_{m(F-6-P)}$ than did PGI-SS. As did the $K_{m(G-6-P)}$ values, the $K_{m(F-6-P)}$ values for PGI-FF and PGI-SS increasingly diverged with increasing temperature. The differences between the allozymes were statistically significant at 20–35 C but not at 10 C, the lowest temperature studied (table 1). As with the $K_{m(G-6-P)}$ values, the $K_{m(F-6-P)}$ values of the PGI heterozygote tended to be outside the range shown by the heterozygotes and closer to the $K_m$ for PGI-SS.

Maximal Velocities

Maximal velocities for the forward and reverse directions exhibited by the three PGI allozymes at 10–35 C are given in table 3. Arrhenius plots of these data (fig. 5) indicate similar temperature-$V_{\text{max}}$ profiles for the allozymes in both G-6-P $\rightarrow$ F-6-P and F-6-P $\rightarrow$ G-6-P.

Reliability of the Kinetic Data

An important advantage in studying the steady-state kinetics of a reversible reaction is that, if $V_{\text{max}}$ and $K_m$ values can be measured in both directions, the reliability of these kinetic parameters can be ascertained by substitution of the $V_{\text{max}}$ and $K_m$ values into the Haldane relation:

$$K_{eq} = \frac{V_{\text{max}}(f) \cdot K_m(r)}{V_{\text{max}}(r) \cdot K_m(f)}.$$  (1)

This yields a value for the equilibrium constant ($K_{eq}$) of the reaction that can then be compared to an independently determined value for $K_{eq}$. Substitution of the experimentally determined $V_{\text{max}}$ and $K_m$ values for the forward and reverse directions obtained in this study into the Haldane relation (eq. [1]) yielded $K_{eq}$ values that closely approximated independently estimated $K_{eq}$ values (table 4). In all cases except
Table 3

<table>
<thead>
<tr>
<th>REACTION AND TEMPERATURE (°C)</th>
<th>SS</th>
<th>FS</th>
<th>FF</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-6-P → F-6-P:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>1424 (7), 10.8</td>
<td>60 (1.5), 11.3</td>
<td>953 (24), 14.4</td>
</tr>
<tr>
<td>30</td>
<td>947 (11), 14.1</td>
<td>39 (0.37), 11.9</td>
<td>598 (6), 14.2</td>
</tr>
<tr>
<td>20</td>
<td>470 (4), 12.3</td>
<td>19 (0.16), 14.8</td>
<td>289 (3), 14.1</td>
</tr>
<tr>
<td>10</td>
<td>271 (4), 19.7</td>
<td>ND</td>
<td>150 (2), 18.5</td>
</tr>
<tr>
<td>F-6-P → G-6-P:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>1339 (26), 11.7</td>
<td>54.49 (0.57), 15.8</td>
<td>819 (10), 15.9</td>
</tr>
<tr>
<td>30</td>
<td>967 (12), 15.8</td>
<td>37.60 (0.57), 19.2</td>
<td>629 (7), 11.7</td>
</tr>
<tr>
<td>20</td>
<td>568 (4), 14.0</td>
<td>ND</td>
<td>388 (5), 22.5</td>
</tr>
<tr>
<td>10</td>
<td>350 (3), 22.4</td>
<td>13.82 (0.11), 14.4</td>
<td>226 (4), 17.0</td>
</tr>
</tbody>
</table>

Note.—ND = No data.

*a* Expressed as μmole/min/ml × 10². Values are lower for FS because rates were measured using a more dilute stock solution.

*b* Fractional values are the consequence of the biweighting procedure (see Material and Methods).

PGI-FF at 35 °C, *Kₚₑₑ* determined from the Haldane relationship was within 10%—and often within 5%—of the experimentally determined *Kₑₑ*. The most deviant *Kₑₑ* value from the Haldane relationship—that for the PGI-FF allozyme at 35 °C—still exhibited a good fit to the expected *Kₑₑ* (table 3). This indicates that the experimentally determined kinetic constants are internally consistent and appear to be reliable estimates of the true kinetic constants.

Fig. 5.—Arrhenius plots of *Vₘₐₓ* for PGI-SS (○) and PGI-FF (●) determined in (A) F-6-P → G-6-P and (B) G-6-P → F-6-P reactions. To facilitate comparisons between the allozymes, *Vₘₐₓ* values were standardized to a common value at 20 °C. K = Degrees Kelvin.
Table 4

\(K_{eq}\) Values Determined by Substitution of Estimated Kinetic Parameters into the Haldane Equation, versus Experimentally Determined \(K_{eq}\) Values

<table>
<thead>
<tr>
<th>Temperature (°C)</th>
<th>(K_{eq}) Value of Allozyme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SS</td>
</tr>
<tr>
<td>35</td>
<td>0.34, 0.32</td>
</tr>
<tr>
<td>30</td>
<td>0.29, 0.30</td>
</tr>
<tr>
<td>20</td>
<td>0.23, 0.25</td>
</tr>
<tr>
<td>10</td>
<td>0.22, 0.20</td>
</tr>
</tbody>
</table>

Note.—For each entry, the first number is the value as determined by substitution of estimated parameters into the Haldane eq. and the second number is the experimentally determined value. Experimentally determined \(K_{eq}\)’s are from Hall (1983), obtained under assay conditions identical to those used in the present study. Ellipses indicate that parameters were not measured in either direction.

Specific Activities

Specific activities for the \(Pgi\) genotypes from each sex and morph type are presented in table 2. In general, within each sex and morph, the \(Pgi^{SS}\) genotype tended to exhibit a slightly higher specific activity than did \(Pgi^{FF}\), with the heterozygote exhibiting intermediate values. However, these differences were not significant as determined by a three-way ANOVA (table 2). A three-way ANOVA of the raw activity data also showed nonsignificant differences, as did separate analyses of genotype-dependent activity in each sex and morph.

\(V_{max}/K_m\) Ratios

In both \(F-6-P \rightarrow G-6-P\) and \(G-6-P \rightarrow F-6-P\), \(V_{max}/K_m\) values were very similar for PGI-FF and PGI-SS at low (10–20°C) temperatures but increasingly diverged at higher temperatures (table 5). Similar results were obtained whether \(V_{max}\) values were standardized at 10 or 20°C (see Material and Methods), and only the values obtained when \(V_{max}\) values were standardized at the latter temperature are reported here. At higher temperatures, PGI-FF in all cases exhibited higher \(V_{max}/K_m\) values than did PGI-SS.

Thermostability

Results of the thermostability studies are given in figure 6. PGI from \(L. canaliculatus\) is fairly labile at temperatures >35°C, with a total loss in activity after incubation for 5 min at 50°C in the absence of 6-PG and >85% loss in activity when incubated at the same temperature in the presence of a saturating concentration of 6-PG. 6-PG substantially increased the resistance of PGI to thermal denaturation in the temperature range of 35–48°C; for example, PGI from each of the three genotypes lost >93% activity when heated at 45°C for 5 min without 6-PG but only lost 9%–12% activity when heated for the same period of time in the presence of 6-PG.

The PGI allozymes exhibited substantial differences in thermostability in the absence of 6-PG (fig. 6). PGI-SS and PGI-FS were virtually indistinguishable in stability characteristics (data for PGI-FS not shown), while the PGI-FF allozyme proved to be much more labile at higher temperatures; for example, during the 5-min incubation period at 40.0°C, PGI-FF lost 32% more activity than did PGI-SS. The difference in
loss of activity between these two allozymes was even greater at 42.5°C. However, in the presence of 6-PG, there were no significant thermostability differences among the allozymes (fig. 6).

The thermal lability of PGI at 35°C in the absence of 6-PG could potentially present problems in the kinetic studies, especially for the more thermally labile PGI-FF enzyme. Consequently, at 35°C rates were measured at saturating and subsaturating F-6-P concentrations at 0–2 as well as at 2–4 min after the addition of the enzyme. This was done in order to determine the extent of any thermal denaturation under conditions employed during the kinetic studies. No differences were observed in the rates for either PGI-FF or PGI-SS, indicating that there was no thermal denaturation during the initial velocity measurements. The absence of thermal denaturation at 35°C in the presence of F-6-P suggests that this compound also stabilizes the enzyme at higher temperatures, an effect similar to that observed with 6-PG.

Discussion

PGI allozymes from *Limnoporus canaliculatus* differ in several important enzymatic characteristics in a manner consistent with expectations of adaptive kinetic differentiation. The $K_m$ values for F-6-P exhibit increasing divergence among the allozymes with increasing temperature (table 1). Moreover, allozyme-dependent variation in the $K_m$ values is inversely related to habitat temperature; that is, the $K_m$ values of the hypothetically warm-adapted PGI-FF enzyme are significantly lower than those of the PGI-SS enzyme at higher temperatures. An identical pattern of allozyme-dependent variation for the $K_m$ values for G-6-P was also observed (table 1).

The overall allozyme-dependent variation of $K_m$ with temperature observed in this study is similar to observed variation among $K_m$ values of the extensively studied lactate dehydrogenase enzymes from species inhabiting different thermal regimes (So-
FIG. 6.—Loss of activity of PGI allozymes after incubation for 5 min at various temperatures in the presence (◦ = PGI-SS; △ = PGI-FF) or absence (● = PGI-SS; ▲ = PGI-FF) of 6-PG. Bars represent SEMs; for points without SEMs, symbols were larger than the SEM bars. Results of t-tests comparing the difference in the loss of activity between PGI allozymes at each temperature: in the absence of 6-PG, % loss in activity of PGI-FF > PGI-SS at 35-47.5°C (P < 0.05 and usually <0.001) and nonsignificant (P > 0.01) at 30 and 50°C; in the presence of 6-PG, results of all t-tests were nonsignificant.

Somero 1978; Graves and Somero 1982). Somero and colleagues have interpreted the inverse relationship between $K_m$ and habitat temperature as the outcome of biochemical adaptation in which the $K_m$ values of the enzyme homologues are "adjusted" to substrate levels occurring at the respective habitat temperatures. This tracking of substrate levels by the $K_m$ values is hypothesized to provide catalytic and regulatory flexibility for enzyme homologues occurring in different thermal habitats. Alternatively, this pattern is consistent with expectations of adaptive catalytic-efficiency differences among the enzymes. The $K_m$ occurs in the denominator of catalytic-efficiency equations (both overall catalytic efficiency, $y^{-1}$, and $k_{cat}/K_m$, the most important component of catalytic efficiency; Albery and Knowles 1976; also see discussions in Hall and Koehn 1983; Hall 1985a, 1985b). Consequently, an inverse relationship between $K_m$ and habitat temperature will contribute to a positive association between habitat temperature and catalytic efficiency.

At this point a caveat must be given. Since PGI from *L. canaliculatus* could not be purified to homogeneity (see Material and Methods), $k_{cat}$ estimates could not be obtained for the allozymes. The kinetic contribution to the reaction rate under physiological conditions is in large part determined by $k_{cat}/K_m$. Thus, it is possible that $k_{cat}$ differences could occur in a direction opposite to and offsetting that portion of the variation in $k_{cat}/K_m$ that is due to allozyme-dependent differences in the $K_m$ values.
In principal, this could result in the abolition or even rank-order reversal of the apparent temperature adaptation of the allozymes that has been inferred on the basis of the $K_m$ data alone. Ultimately, this issue can only be settled by future kinetic studies of PGI allozymes from *L. canaliculatus*, studies in which $k_{cat}$ estimates are obtained on homogeneously purified allozymes.

Note that inferences of kinetic adaptation drawn from $K_m$ data alone often agree well with those drawn from $k_{cat}/K_m$ data. For example, studies by Place and Powers (1979), Graves and Somero (1982), and Hall (1985b) each reported a negative association between $K_m$ and habitat temperature—a finding similar to that of the present study—and a positive association between $k_{cat}/K_m$ and habitat temperature. In none of the aforementioned studies did variation in $k_{cat}$ abolish or even significantly diminish the contribution of $K_m$ to variation in $k_{cat}/K_m$. In all cases, variation in $K_m$ was either the major contributor to variation in $k_{cat}/K_m$ among the enzymes (Place and Powers 1979; Graves and Somero 1982) or variation in $k_{cat}$ contributed as much did variation in $K_m$ (Hall 1983). Thus, in each of these three cases, inferences of kinetic adaptation based solely on the $K_m$ values are identical to those based on $k_{cat}/K_m$ values. Similar results were also obtained for esterase-6 allozymes (Mane et al. 1983), in which variation in $k_{cat}/K_m$ (measured at one temperature) for $\beta$-naphthylpropionate was due almost exclusively to variation in $K_m$.

$V_{max}/K_m$ ratios also varied between Pgi genotypes in a manner consistent with expectations of adaptive differentiation (table 5). At higher temperatures, the hypothetically warm-adapted PGI-FF enzyme was associated with a higher $V_{max}/K_m$ ratio than was the PGI-SS enzyme. An enzyme preparation possessing a higher $V_{max}/K_m$ ratio under a specific set of conditions is catalytically more efficient than one possessing a lower $V_{max}/K_m$ ratio; however, unlike variation in $k_{cat}/K_m$—which, by definition, must be due to allozyme-dependent kinetic variation—it is often not possible to specify whether variation in $V_{max}/K_m$ is a consequence of allozyme-dependent kinetic variation or of allozyme-associated regulatory variation. This is so because $V_{max}$ is a composite parameter determined by two independent parameters, $k_{cat}$ and enzyme concentration. It is therefore possible that variation in $V_{max}/K_m$ ratios among allozymes could be due entirely to variation in enzyme concentration as a consequence of variation at a tightly linked regulatory locus. This important distinction has not been appreciated by several workers who have compared $V_{max}/K_m$ ratios of allozymes (Watt 1977, 1983; Hoffman 1981a). However, the ambiguity as to whether variation in $V_{max}/K_m$ is a consequence of kinetic or regulatory variation is only a problem when variation in $V_{max}$ is a major contributor to differences in $V_{max}/K_m$. In the case of PGI allozymes from *L. canaliculatus*, $V_{max}$ values do not differ among PGI genotypes (table 2) and the PGI genotypes exhibit essentially identical $V_{max}$-temperature profiles (fig. 5). This means either that $k_{cat}$ values do not differ among the allozymes or that $k_{cat}$ values differ but are canceled out by reciprocal variation in enzyme concentration between the allozymes that results in identical $V_{max}$ values. For whatever reason, the absence of differences in $V_{max}$ among the allozymes indicates that the variation in $V_{max}/K_m$ ratios is due to allozyme-dependent kinetic variation (i.e., variation in $K_m$ values).

In summary, the kinetic data, taken together, indicate that the PGI-FF allozyme is kinetically superior to the PGI-SS allozyme at higher temperatures. However, the biochemical data only explain part of the PGI latitudinal variation—namely, why $Pgi^F$ should increase in southern latitudes. There is no biochemical explanation for the increase in the frequency of $Pgi^S$ in northern latitudes, since there was no case in
which the PGI-SS allozyme was kinetically superior to the PGI-FF allozyme at lower temperatures.

Significance of the allozyme-dependent differences in thermostability is more difficult to ascertain. Although the allozymes exhibited substantial differences in thermal stability at temperatures likely to be encountered by *L. canaliculatus* in the field (see Material and Methods), these differences were contingent on the absence of 6-PG in the incubation medium (fig. 6). The strong stabilizing influence of 6-PG (and probably other substrates and substrate analogues; see Results) means that the expression of allozymic differences in thermal stability in vivo will be determined by the relative proportion of unbound enzyme vis-à-vis that bound to substrates and substrate analogues. Consequently, the biological significance of the in vitro allozymic differences in thermal stability can only be determined by future in vivo studies of this characteristic.

If the allozymic differences in thermal stability are expressed in vivo, then this property appears to be maladaptive, at least in the portion of the cline north of Florida. In this area, PGI-FF increases in southern latitudes yet is more thermolabile than is PGI-SS. It is possible that thermostability is a more important enzymatic characteristic in the continuously warm subtropical habitats of southern Florida, which might account for the reversal of the PGI cline in this area.

One unusual feature of this study was the tendency for the heterozygote to be outside the range of the homozygotes in a variety of properties. This was generally observed for the $K_m$ values for both substrates (table 1), as well as for the inhibition constants ($K_i$ values) for 6-PG (Zera 1984, and accepted). Since the $K_m$ values and the $K_i$ values for the heterozygote were higher than those for the homozygotes, the overdominance is negative. Since negative overdominance resulting from heterozygote inferiority cannot lead to a stable polymorphism, the overdominance observed in this study cannot account for the persistence of the $Pgi^S$ allele in the absence of any observed kinetic advantage of the PGI-SS enzyme over the PGI-FF enzyme.

Although the majority of enzyme polymorphisms result in the intermediacy of kinetic properties of the heterozygote (Koehn et al. 1983; Zera et al. 1985), this is not universally observed (Martin 1979; McKetchnie et al. 1981; Watt 1983). Presumably, these examples of overdominance are the result of nonadditive effects due to subunit interactions in the heterodimer. At present there are no studies that have isolated and kinetically characterized the heterodimer of a polymorphic enzyme.

Besides the present study, there are two other biochemical characterizations of clinally varying PGI allozymes (Hoffman 1981a, 1983; Hall 1985b). How do the results of these studies compare with each other, and do any general patterns emerge? Hall (1985b) estimated, at several temperatures and under physiological conditions, all the component parameters of catalytic efficiency for the two common latitudinally varying PGI allozymes from the bivalve mollusc, *Mytilus edulis*. He observed significant differences in overall catalytic efficiency ($y^{-1}$) and in $k_{cat}/K_m$ ratios among the PGI allozymes at higher (15–25 C) but not at lower (5–10 C) temperatures. The differences in both of these parameters were in accordance with expectations of adaptive kinetic differentiation based on the latitudinal Pgi gene-frequency cline.

In the other kinetic study of latitudinally varying PGI allozymes, this one in the sea anemone, *Metridium senile*, Hoffman (1981a) observed kinetic variation among the allozymes in the $K_m$ values for G-6-P and F-6-P. In this case, it is somewhat difficult to draw strong inferences of adaptation from the data. The rank order of
allozymic differences in $K_m$ values is reversed in the forward versus the reverse reaction directions, and the kinetic studies in the forward (G-6-P $\rightarrow$ F-6-P) direction were performed at a nonphysiological pH (pH 8.5 at all temperatures). Nevertheless, the kinetic data suggest that the PGI-FF allozyme is catalytically superior to the PGI-SS allozyme at higher temperatures. In addition, the results are consistent with the gene-frequency data, in which the $Pgi^F$ allele increases in southern latitudes (Hoffman 1981b).

The correspondence, in each of these three species, between the kinetic properties of the PGI allozymes and the latitudinal variation of the PGI allele frequencies implicates selection as having some influence on these PGI polymorphisms. Additional studies on the physiological effects of the allozymic variation as well as direct measurements of fitness differences among the enzyme genotypes are required to substantiate the role of selection.

Although these data suggest that natural selection is at least partially responsible for the PGI clinal variation in the three aforementioned species, the data are equivocal as to whether the clines are stable or transient. As mentioned above, the work of Hoffman (1981a, 1983) and Hall (1985b), as well as the present study, have documented only that PGI from one genotype is catalytically superior to that from the other genotype(s), primarily or solely at higher temperatures. Hoffman (1981a) has reported preliminary evidence on the differential inhibition of PGI allozymes of Metridium by the pentose-shunt metabolite, 6-PG, at low temperatures. This difference could provide a selective advantage for the $Pgi^{SS}$ genotype at colder temperatures (see Discussion in Hoffman 1981a) and thus could provide the basis for the maintenance of the PGI cline in Metridium via balancing selection. On the other hand, detailed studies of the inhibition of PGI allozymes of Mytilus (Hall 1985b) and Limnoporus (Zera 1984, and accepted) by 6-PG and other pentose-shunt metabolites at several temperatures failed to document any differences among the allozymes. In short, at present there is no strong evidence supporting the existence of stable PGI clines maintained by balancing selection in Metridium, Mytilus, or Limnoporus. It is therefore possible that the PGI clines, while apparently influenced by selection, are nevertheless transient and that one allele will eventually reach fixation in each species.

Acknowledgments

The computer program used to estimate kinetic constants was kindly provided by A. Cornish-Bowden. J. Lo Dolce and F. J. Rolf helped interface the microcomputer with the spectrophotometer and wrote programs for the collection of initial rate data. I wish to thank J. G. Hall for advice on enzyme kinetics and enzyme purification, M. E. Saks for thoughtful comments on various drafts of the manuscript, and M. Nei for pointing out ambiguities in the presentation of the $Pgi$ allele frequency cline in an earlier draft of this manuscript. I especially thank R. K. Koehn for advice, encouragement, and financial support during the course of this study. This work was supported by National Science Foundation grant DEB 7908802 and U.S. Public Health Service grant GM 21133 to R. K. Koehn and a National Science Foundation dissertation-improvement grant to A.J.Z. This is a portion of a dissertation submitted to the graduate school in partial fulfillment of the Ph.D. degree in biological sciences and is contribution 601 in Ecology and Evolution at the State University of New York at Stony Brook.
LITERATURE CITED


HALL, J. G. 1983. Temperature-related catalytic differentiation among homologues of glucose-phosphate isomerase (EC 5.3.1.9) isolated from the bivalve molluscs, Mytilus edulis and Isognomon alatus. Ph.D diss., State University of New York at Stony Brook.


MCKETCHNIE, S. W., M. KOEHN, and S. C. PHILLIPS. 1981. A search for interacting polymorphic


**MASATOSHI NEI, reviewing editor**

Received May 27, 1986; revision received October 20, 1986.
Evolutionary Diversification of Class II P Loci in the Mhc of the Mole-Rat *Spalax ehrenbergi*

Ralf Schöpfer,* 2  Felipe Figueroa,* Dean Nižetić,* Eviatar Nevo,† and Jan Klein*

*Max-Planck-Institut für Biologie, Abteilung Immungenetik; and †Institute of Evolution, University of Haifa

The class II region of the major histocompatibility complex (*Smh*) in the mole rat, *Spalax ehrenbergi*, consists of only two gene families, *P* and *Q*, instead of the four families (*P*, *O*, *Q*, and *R*) found in all other mammals studied to date. The *Spalax* *P* family consists of at least four *β* and three *α* genes or gene fragments. In DNA-hybridization experiments, two of the *α* genes behave as bona fide *P*-family members in that they hybridize strongly with human DPβ probes and hybridize weakly with probes specific for other class II gene families. The other two *β* genes, on the other hand, hybridize weakly with human DPβ probes and nearly as well with human DQα probes. To determine the evolutionary relationships among these *P*-like genes, we have sequenced one of them. The sequence reveals, on the basis of its overall organization, that the gene clearly belongs to the *P* family, yet, on the basis of its nucleotide sequence, it is only slightly more similar to human DP than to human DQ genes. These results indicate that in the *Spalax* the *P* family of genes split into two subfamilies, *PA* and *PB*. For unknown reasons, one of these subfamilies (*PB*) retained more similarity to the *Q* gene family than did the other (*PA*).

Introduction

The major histocompatibility complex (Mhc) is a cluster of loci involved in the recognition of foreign antigens by T-lymphocytes of the host immune system (see review in Klein 1986). The number of loci and their organization vary considerably from species to species (Klein and Figueroa 1986). Some species, such as the pig, may contain as few as 20 loci in their Mhc (Chardon et al. 1985), while others, such as the mole rat, contain >65 loci (V. Vincek, D. Nižetić, M. Golubić, F. Figueroa, E. Nevo, and J. Klein, unpublished data). All the Mhcs identified so far consist of two classes of loci involved in the recognition of different kinds of antigen. Both class I and class II loci can be divided into two subclasses, α and β, and into several families on the basis of (1) the degree of relatedness and, presumably, (2) their evolutionary histories (fig. 1). At least two families of class I loci and five families of class II loci are known to exist. The Mhc families have different designations in different species. The presumed correspondence among the class II families of the human, mouse, and rat are shown in table 1. Most of the class I genes appear to be nonfunctional pseudogenes (Klein et al. 1983). Some of the functional class I and class II loci are highly polymorphic;

1. Key words: major histocompatibility complex, Mhc, DNA sequencing, mole rat, *Spalax ehrenbergi*, evolution.

2. Current address: The Salk Institute for Biological Studies.

Address for correspondence and reprints: Dr. Jan Klein, Max-Planck-Institut für Biologie, Abteilung Immungenetik, Corrensstraße 42, 7400 Tübingen 1, Federal Republic of Germany.

© 1987 by The University of Chicago. All rights reserved.
0737-4038/87/0403-0007$02.00
more than 100 alleles exist at some of the loci in some species (Klein and Figueroa 1981).

The complexity, polymorphism, and high frequency of nonfunctional loci in the Mhc raise many questions concerning the evolution of this cluster. The information about the evolution of the Mhc is sketchy, however, since only the Mhcs of two species, the human and the mouse, have been studied in detail. Furthermore, since the organization of the human Mhc is very different from that of the mouse Mhc, it is difficult to draw firm conclusions about the origin of the individual loci. In particular,

Table 1

<table>
<thead>
<tr>
<th>FAMILY</th>
<th>Human (HLA)*</th>
<th>Mouse (H-2)</th>
<th>Rat (RT1)</th>
<th>Mole Rat (Smh)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>DP_\alpha</td>
<td></td>
<td></td>
<td>P_\alpha</td>
</tr>
<tr>
<td></td>
<td>DP_\beta</td>
<td>A_\beta_3</td>
<td></td>
<td>P_\beta</td>
</tr>
<tr>
<td>Z</td>
<td>DZ_a</td>
<td>A_\beta_3</td>
<td>Absent</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>DO_\beta</td>
<td>A_\beta_2</td>
<td>Absent</td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>DQ_a</td>
<td>A_\gamma_1</td>
<td>B_\gamma</td>
<td>Q_\alpha</td>
</tr>
<tr>
<td></td>
<td>DQ_\beta</td>
<td>A_\gamma_1</td>
<td>B_\gamma</td>
<td>Q_\beta</td>
</tr>
<tr>
<td>R</td>
<td>DR_a</td>
<td>E_\gamma_1</td>
<td>D_\gamma</td>
<td>Absent</td>
</tr>
<tr>
<td></td>
<td>DR_\beta</td>
<td>E_\gamma_1E_\gamma_2</td>
<td>D_\gamma</td>
<td>Absent</td>
</tr>
</tbody>
</table>

*Individual genes in each family are numbered (i.e., DP_\alpha_1, DP_\alpha_2, DP_\beta_1, DP_\beta_2, and so on). Note, however, that A_\beta_2 and A_\beta_3 are not members of the A family; here, the letters and numbers have no relation to the family membership. (This nomenclature may seem irrational, which it is, but we did not invent it.)
it is important to ascertain whether the Mhc organization found in the mouse is representative of that of all rodents. As a step toward answering this question, we have initiated the study of the Mhc in the mole rat, *Spalax ehrenbergi* (Nižetić et al. 1984, 1985, and submitted). The mouse and the mole rat are both myomorph rodents, but their estimated time of separation is 20–40 Myr before the present (MYBP). The mole rat is thus close enough to the mouse to provide information about the events that preceded the generation of the mouse-type Mhc organization—and, at the same time, it is different enough from the mouse to reveal how general this organization is.

In an earlier publication, we described the organization of the class II loci of *S. ehrenbergi* (Nižetić et al., submitted). As it turns out, this organization is very different from that of the mouse. Two of the four families (*O* and *R*) are missing in the mole rat, whereas a third family (*P*) has been greatly expanded in comparison with that of the mouse. The function of the human *DR* family has apparently been taken over by the *P* family in the mole rat. Furthermore, we noticed that some of the *Spalax* class II loci were difficult to assign to a particular family on the basis of DNA hybridization with probes specific for human or mouse families. Specifically, some of the mole-rat loci appeared to hybridize weakly with both *P*- and *Q*-specific probes and virtually not at all with *R*- or *O*-specific probes. The question therefore arose whether these loci represent a new, as yet unidentified, family of class II loci or whether their behavior may not reflect some peculiarity in the evolution of class II loci. To provide an answer to this question, we have sequenced the DNA constituting one of these loci. The results are described below.

**Material and Methods**

**Isolation and Characterization of Clones**

A genomic library, constructed from the kidneys of a single adult female mole rat, *Spalax ehrenbergi*, belonging to the *2n = 58* chromosome species (see Nevo 1985), was screened with mouse *A*, and *A* class II probes (Nižetić et al., submitted). The cosmids clone 8.4 was found to contain two β genes and one α gene fragment, which we designate *P*₁₁, *P*₂₂, and *P*ₐ₂, respectively. A second cosmids clone, 9.6, was discovered to contain the *P*₁₁ and *P*₂₂ genes, part of the *P*₂₂ gene, as well as another α gene that we designate *P*ₐ₁. The segments derived from cosmids clones 8.4 or 9.6 and containing the *P*₂₂ gene were then subcloned in plasmid vectors (fig. 2). A 750-bp *BamHI* fragment was isolated from the cosmids 9.6 and subcloned in *pUC8* (Vieira and Messing 1982) to give the subclone B750. All the other subclones were derived from the clone 8.4. The 9-kb *EcoRI*, the 8-kb *EcoRI/Sall*, and the 6-kb *KpnI* fragments were subcloned and designated 8.4.E4.3, 8.4.ES8.1, and 8.4.K2.5, respectively (the first fragment was subcloned in *pUC8*, the remaining two in *pUC18*). A detailed restriction map of these fragments was then constructed following single or double digestions with appropriate enzymes, and subfragments were selected with exon-specific human DPᵦ probes (fig. 3). The subfragments were then isolated and subcloned in M13mp18 (Yanisch-Perron et al. 1985).

---

3. Data on hemoglobin divergence (Kleinschmidt et al. 1985) and microcomplement fixation (Nevo and Sarich 1974) suggest a separation time of 40 MYBP. By contrast, DNA-DNA hybridization analysis indicates a divergence time of 20 MYBP (F. M. Catzeflis, J. E. Ahlquist, E. Nevo, and C. G. Sibley, unpublished data). The latter estimate is probably more realistic.

4. The mouse is quite distant from *Spalax* in several other genetic systems as well, e.g., myoglobins (Guernett et al. 1984) and liver fructose-1,6-biphosphatase (Rittenhouse et al. 1985).
FIG. 2.—Restriction enzyme map of the Smh-PBβ gene (1) and DNA sequencing strategy for this gene (2). Exons (black rectangles) are numbered 1-6. Segments represent cosmids covering the gene. Letters a-d in 2 indicate subclones used for sequencing (their origin is indicated by brackets and corresponding letters in 1); and arrows indicate the lengths of DNA regions that were sequenced. The 3'-labeled ends are at the arrow tails. Enzyme sites: A = AluI; B = BamHI; C = SalI; E = EcoRI; H = HaeIII; Hi = HindIII; K = KpnI; M = SmaI; N = XhoI; P = PstI; S = SacI; T = TaqI; U = Sau3A; and X = XbaI. Other abbreviations (below the lines): A, B, and C = three consensus sequences of the promoter region; SUT = 5'-untranslated region; and LP = leader peptide.

HLA-DPβ Probes

The HLA-DPβ probe pSB2β is a full-length cDNA clone; it is 1,079 bp long and contains all the protein-encoding exons as well as the 5'- and 3'-untranslated regions, including the poly(A) signal (Kappes et al. 1984). Four exon-specific probes were derived from the pSB2β probe: probe I (350 bp), specific for exon 2; probe II (850 bp), encompassing exons 1-5; probe III (250 bp), specific for exon 6; and probe IV (650 bp), encompassing exons 3-6 (see fig. 3).

Nucleotide Sequence Analysis

Selected fragments were subcloned in M13mp18 and sequenced using the 2'3' dideoxynucleotide chain-termination method of Sanger and Coulson (1975) and Sanger et al. (1977) with α32PdATP (New England Nuclear, Dreieich, Federal Republic of Germany; see Biggin et al. 1983). The sequencing strategy is depicted in figure 2. Both DNA strands were sequenced.
Other Material and Methods

Southern blotting, hybridizations, and digestions were performed—and large-scale preparations of cloned DNA were made ready—as described previously (Nižetić et al., submitted). Restriction enzymes were purchased from Pharmacia (Freiburg, Federal Republic of Germany) or Boehringer (Mannheim, Federal Republic of Germany). Klenow polymerase was also from Boehringer Mannheim.

Results

The sequence of all the putative exons and parts of the introns of the *Spalax* gene was determined by the chain-termination method, using the strategy outlined in figure 2. The nucleotide sequence and the amino acid sequence derived from it are shown in figure 4. The putative location of the exons was determined by comparison with published sequences of class II genes, in particular the human DPβ1 gene (Kelly and Trowsdale 1985). The *Spalax* gene spans some 12 kb, of which we have sequenced 3.4 kb. The predicted size of the messenger RNA is 1,150 bp.

Alignment of the *Spalax* sequence upstream from the ATG initiation codon with mouse and human class II sequences reveals three highly conserved segments that most probably represent the promoter region of the *Spalax* gene (the human DPβ1 gene equivalents are given in parentheses in the following comparisons [Kelly and Trowsdale 1985]). Located farthest upstream is a segment of 21 (21) bp corresponding to the consensus A block of Kelly and Trowsdale (1985). Fifteen (15) nucleotides downstream from it is another conserved sequence (consensus B) of 15 (15) bp. Finally, another 15 (17) bp downstream is the CAT sequence (CCAATCC), and the initiation codon (ATG) is then 112 (120) bp downstream from the CAT sequence. The start site is postulated to be 43 (51) bp downstream from the CAT sequence. The distances between these individual elements of the promoter region are very similar to those found in other class II genes, and they correspond almost precisely to those present in the human DPβ genes.

Exon 1 of the *Spalax* gene codes for the 5'-untranslated region, the leader sequence, and the first five amino acids of the first extracellular domain of the β chain (the B1 domain). The postulated leader sequence is of the same length (29 amino acid residues) as that in the human DPβ1 gene. (In most other class II genes it is of a slightly different
Evolution of Mhc 293

The two leader sequences differ in only eight of the 29 amino acid residues. As in other leader sequences, the amino acids are mostly of the hydrophobic type, although the Spalax sequence does contain one charged amino acid (Lys at position 6).

Exon 2 codes for the rest of the β1 domain, specifically for residues 6–93. This length of the β chain is two residues shorter than that in most other class II β chains sequenced thus far, with the exception of human DPβ chains. The shortening can be explained by a deletion of two residues (two codons in the DNA sequence) that correspond to residues 23 and 24 in the DR or DQ β chains. The Spalax and the human DPβ genes both have this deletion in exactly the same place.

The Spalax sequence contains two cysteines in the same position as in the other class II β chains (15 and 77). These two residues are presumably connected in the folded chain by a disulfide bond. All functional human and mouse class II β chains described thus far have an N-glycosylation site at position 19 (Asn-Gly-Thr in the human DPβ1 gene), but in the Spalax gene this site is lacking (the corresponding sequence is His-Gly-Arg).

Exon 4 codes for the so-called connecting peptide, the transmembrane region, and five amino acid residues of the cytoplasmic tail (the exon codes for residues 188–224). The transmembrane region consists of 22 amino acid residues, most of which are hydrophobic, the rest being uncharged.

Exon 5 codes for the remaining six amino acid residues of the cytoplasmic tail (residues 225–229) and is followed by a stop codon TAA and then by four nucleotides of the 3'-untranslated region.

Exon 6 codes for the rest of the 3'-untranslated region. It is marked downstream by an atypical polyadenylation signal AATAAG (see Birnstiel et al. 1985), followed by a GT-rich region. No other polyadenylation signal could be found in the stretch of the DNA sequenced. If the AATAAG signal is indeed the one used, exon 6 would be ~310 bp long and would end somewhere in the GT-rich region. A similar situation again exists in the human DPβ1 gene, in which the polyadenylation signal is AATAAT and exon 6 is 229 bp long (Kelly and Trowsdale 1985).

The Spalax gene also resembles the human DPβ genes in the length of its introns. Characteristic features of the DPβ genes are the long introns (almost 5 kb) separating exon 1 from exon 2 and exon 2 from exon 3. In the Spalax gene, the lengths of these two introns are 4.9 kb and 4.7 kb, respectively. The remaining introns are short: introns 3, 4, and 5 are 480 bp, 390 bp, and 250 bp long, respectively. Theoretically, the joining of exons 4 and 5(6) can occur in several ways (Some of these are shown in fig. 5). This is possible because at the 3' end of exon 4 there are several GT splicing

FIG. 4.—Nucleotide sequence of the Smh-PBβ gene. Exons are underlined; and elements of the promoter region and the polyadenylation signal are boxed. The deduced amino acid sequences are also indicated. Stop codons are indicated by triple asterisks. The dots in the introns indicate that a sequence of unspecified length has not been determined.
sites. The alternative splicings could produce cytoplasmic regions of different length and composition. Which of these sites are actually used is not clear, but in humans, where a similar situation exists at the $DP_{\beta_1}$ locus, at least two splicing variants have been found (Kelly and Trowsdale 1985).

**Discussion**

The sequenced *Spalax* gene is obviously a member of the $P\beta$ family of class II genes to which the previously identified human $DP_{\beta}$ (Kelly and Trowsdale 1985) and the mouse $A_{\beta_3}$ (Widera and Flavell 1985) genes belong. This conclusion is based on several observations. First, the *Spalax* gene contains two very long introns, which are approximately of the same length (almost 5 kb) as those separating exon 1 from exon 2 and exon 2 from exon 3 in the human $DP_{\beta}$ genes. Such long introns have thus far not been found in any other family of class II genes. Second, the *Spalax* gene and the human $DP_{\beta}$ genes have a very similar organization of their promoter regions. The distances between the regulatory elements in the putative promoter region, as well as the length of the elements themselves, correspond precisely in the *Spalax* and the human $DP_{\beta}$ genes. Third, the nucleotides as well as the amino acid sequence of the leader segment in the *Spalax* $\beta$ gene is much more similar to that of the human $DP_{\beta}$ genes than it is to the genes in other class II families. The same applies to the length of the leader sequence. Fourth, the overall exon-intron organization is the same in the *Smh* $\beta$ and the *HLA-DP_{\beta}* genes. Fifth, similar options for alternative splicing at the 3' end (between exons 4, 5, and 6) exist in the *Spalax* $\beta$ gene and in the human $DP_{\beta}$ genes. Sixth, following amino acid residue number 22 in the $\beta_1$ domain, two residues have been deleted in the *Spalax* $\beta$ gene when compared with $\beta$ chains belonging to the other class II families. The same two residues are also lacking in the $DP_{\beta}$ genes. (Although two codons have been deleted in that general region in one of the human $DR_{\beta}$ pseudogenes, only one of these two overlaps with the codons deleted in the *HLA-DP_{\beta}* and *Smh* $\beta$ genes. Furthermore, in the $DR_{\beta}$ pseudogene, the deletion has occurred in the vicinity of a stop codon and is clearly an event independent of the deletion in the $DP_{\beta}$ genes; see Larhammar et al. 1985.) Seventh, the *Smh* $\beta$ chain contains a number of amino acid residues that are also present at the same position in the human $DP_{\beta}$ genes but are absent in all the other known class II $\beta$ genes (table 2). These residues can therefore be considered $P$-family specific. Eighth, the nucleotide sequence of the individual exons of the *Smh* $\beta$ gene is most similar to that of the human $DP_{\beta}$
Evolution of Mhc 295

genes (table 3). Similarly, the overall amino acid sequence of the Smh β chain most closely resembles the sequence of the DPβ chains (not shown). We conclude, therefore, that the Smh β gene is a member of the P family and designate the gene Smh-Pβ1.

Why does the Smh-Pβ1 DNA hybridize nearly as well with human DQβ as with human DPβ probes (as reported by Nižetić et al. [submitted])? The reason is apparently that the Smh-Pβ1 gene also shows high sequence similarity with the human DQβ genes (table 3). Although this similarity is somewhat lower than that to the human DPβ genes, the difference is so small that it apparently cannot be detected by the DNA-hybridization method and the probes that we used. The situation is, then, such that, in its overall organization, the Smh-Pβ1 gene is clearly a member of the P family but, in its sequence similarity, is only slightly more related to the P-family genes than it is to the Q-family genes. The Spalax P family also contains genes that hybridize strongly with human DPβ probes and weakly with human DQβ probes. The family thus contains two subfamilies of genes: one whose members are closely related to human DPβ genes and relatively unrelated to HLA-DQβ genes and another whose members are only slightly more related (in terms of nucleotide sequence) to the DPβ rather than to the DQβ genes. We shall call the former the PA and the latter the PB subfamily of genes.

What could be the explanation for the close relationship of the PB loci to the loci of the Q family? One possibility is that information has repeatedly been transferred during evolution from Q-family genes to the PB-subfamily members by the mechanism of gene conversion. Since the two families are probably located relatively close together on the chromosome, such a transfer might have been possible. However, we consider this possibility unlikely because no blocks of PB/Q-shared sequences, which such a transfer should generate, are found in the Smh-PBβ gene (table 2).

Another possibility is that the PB/Q-gene similarity is the result of convergent evolution. One might speculate that after the P and Q gene families separated from each other, and after they had evolved independently for some time, selection pressure acted on some of the P-family genes to make them fix mutations similar to those that were being fixed in the Q-family genes. Because it is contradicted by the sequence data, this explanation, too, is unlikely. If the observed similarity were the result of convergent evolution, the selection would primarily be for similarity of the amino acid sequence and not of the nucleotide sequence. Yet the similarity between the Smh-PBβ gene and the HLA-DQβ genes also extends to the codon level, including third-nucleotide positions.

The third possibility is that the sharing of the sequences reflects a particular evolutionary history of the genes (fig. 6). One could postulate that the P- and Q-family genes were the last to split off from a common stem in the evolution of the class II genes. Later, the P branch separated into two branches, PA and PB, which then evolved at different rates. The PA branch evolved relatively rapidly and so lost much of the P/Q similarity, whereas the PB branch evolved slowly and thus retained more of the P/Q similarity. The difference in the evolutionary rates between the PA and PB branches could have been caused by selection. In the past, the PA family might have been the most functionally active group of class II genes, particularly in the ancestors of Spalax, in which the R family of genes has been deleted (Nižetić et al., submitted). The PB subfamily, on the other hand, might have been under less selection pressure. Supporting this contention is the observation that human genes representing this subfamily have not been found thus far, suggesting that, in primates at least, this evolutionary line of class II genes is in a regression phase. (PB-like genes, however, may be present in the rabbit; see Sittisombut and Knight [1986].)
Table 2
Comparison of Amino Acid Residues at Selected Positions in the Class II β Chains of the Mouse (H-2), Man (HLA), Rat (RT1), and Spalax (Smh)

<table>
<thead>
<tr>
<th>POLYPEPTIDE CHAIN</th>
<th>7</th>
<th>13</th>
<th>16</th>
<th>18</th>
<th>25</th>
<th>44</th>
<th>54</th>
<th>55</th>
<th>56</th>
<th>57</th>
<th>58</th>
<th>73</th>
<th>98</th>
<th>105</th>
<th>106</th>
<th>108</th>
<th>124</th>
<th>126</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-2Aβb</td>
<td>F</td>
<td>G</td>
<td>Y</td>
<td>T</td>
<td>V</td>
<td>E</td>
<td>P</td>
<td>D</td>
<td>A</td>
<td>E</td>
<td>Y</td>
<td>L</td>
<td>V</td>
<td>E</td>
<td>A</td>
<td>N</td>
<td>K</td>
<td>K</td>
</tr>
<tr>
<td>H-2Aβd</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H-2Aβk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DQβ1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DQβ2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DQβ3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DQβ4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DQβ5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT1-Åβ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H-2β2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H-2β3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H-2β4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H-2β5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DRβ1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DRβ2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DRβ3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DRβ4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DRβ5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DRβ6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DPβ1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DPβ2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DPβ3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HLA-DPβ4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smh-PBβ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H-2Aβ3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.—References for these sequences (except that of the Smh chain) are to be found in Figueroa and Klein (1986). The IUB one-letter code is used for the amino acids. A dash indicates that the amino acid is the same as in H-2Aβb; an "X" indicates an unknown amino acid; and a dot indicates that the residue has not been determined.

It should be pointed out that the introns of the Smh-PBβ1 gene show no sequence similarity to the human DP genes—or, for that matter, to any other class II genes. This finding is rather puzzling, and we have no good explanation for it. Since we have sequenced only short segments of the long introns, it is of course possible that portions of the introns will turn out to be class II-like when the entire gene is sequenced. If so, the dissimilar sequences could best be thought of as foreign elements that have been inserted into the gene in the Spalax evolutionary line.

Whether the Smh-PBβ1 gene is expressed on the cell surface or not is unclear at the moment. The gene is probably transcribed and translated normally, but the polypeptide chain is probably not glycosylated because the two usual sites of carbohydrate attachment found in expressed class II β chains have been altered. The Smh-PBβ1 chain contains one potential glycosylation site at the border between the connecting peptide and the transmembrane region, but it is unlikely that this site is used—and,
### Table 3

Percent Identity at the DNA Level between the Smh-P<sub>β</sub> Gene and Other Class II β Genes

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA-DP&lt;sub&gt;β&lt;/sub&gt;</td>
<td>74.3 (27)</td>
<td>74.3 (71)</td>
<td>82.3 (50)</td>
<td>84.0 (18)</td>
<td>83.3 (3)</td>
<td>78.6</td>
</tr>
<tr>
<td>RT1-B&lt;sub&gt;β&lt;/sub&gt;</td>
<td>N.A.</td>
<td>71.6 (74)*</td>
<td>75.2 (70)</td>
<td>71.4 (32)</td>
<td>54.2 (8)</td>
<td>72.2</td>
</tr>
<tr>
<td>H-2A&lt;sub&gt;β&lt;/sub&gt;</td>
<td>47.4 (55)</td>
<td>68.5 (87)</td>
<td>76.6 (66)</td>
<td>70.5 (33)</td>
<td>58.3 (7)</td>
<td>68.7</td>
</tr>
<tr>
<td>HLA-DQ&lt;sub&gt;β&lt;/sub&gt;</td>
<td>51.9 (51)</td>
<td>67.0 (91)</td>
<td>78.4 (61)</td>
<td>76.8 (30)</td>
<td>58.3 (7)</td>
<td>69.7</td>
</tr>
<tr>
<td>H-2E&lt;sub&gt;β&lt;/sub&gt;</td>
<td>50.4 (52)</td>
<td>63.4 (101)</td>
<td>72.7 (77)</td>
<td>74.1 (29)</td>
<td>45.8 (9)</td>
<td>66.2</td>
</tr>
<tr>
<td>HLA-DR&lt;sub&gt;β&lt;/sub&gt;</td>
<td>44.9 (58)</td>
<td>63.4 (101)</td>
<td>70.0 (85)</td>
<td>72.3 (31)</td>
<td>50.0 (8)</td>
<td>64.4</td>
</tr>
<tr>
<td>HLA-DO&lt;sub&gt;β&lt;/sub&gt;</td>
<td>44.0 (59)</td>
<td>58.0 (116)</td>
<td>65.6 (97)</td>
<td>66.1 (38)</td>
<td>25.0 (13)</td>
<td>59.2</td>
</tr>
</tbody>
</table>

**NOTE.**—Numbers in parentheses are number of nucleotide differences. Sequence data for non-Smh genes were taken from references listed in Figueroa and Klein (1986). N.A. = Not analyzed; H-2 = mouse Mhc; HLA = human Mhc; and RT1 = rat Mhc.

* 261 nucleotides compared.
FIG. 6.—Postulated evolutionary origin of class II gene families (based on nucleotide sequence comparisons and derived from matrices of pairwise comparisons; for further details, see Klein and Figueroa, 1986).

if it is used, it is improbable that the chain can still be integrated into the membrane. The question of the cell-surface expression of the Smh-PB$_{\beta_1}$ gene will have to be decided by transfection experiments.

Acknowledgments

We thank Dr. Isolde Riede for introducing one of us (R.S.) to the chain-termination sequencing technique and Dr. Jack Strominger (Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Mass.) for the DP$_{\beta}$ probe. E.N. thanks the Israel Discount Bank Chair of Evolutionary Biology and the Ancell-Teicher Foundation for Genetics and Molecular Evolution established by Florence and Theodore Baumritter of New York. This work was supported by a grant from the National Cancer Institute, National Institutes of Health, Bethesda, Md.

LITERATURE CITED


NIŽETIĆ, D., F. FIGUEROA, Z. DEMBIĆ, E. NEVO, and J. KLEIN. Mhc organization in *Spalax ehrenbergi*: evidence for transfer of function between class II genes. Submitted for publication.


WALTER M. FITCH, reviewing editor

Received September 9, 1986; revision received November 12, 1986.
Evolution of the Primary and Secondary Structures of the E1a mRNAs of the Adenovirus

D. A. M. Konings, P. Hogeweg, and B. Hesper
Bioinformatica

In this paper we investigate and compare (evolutionary) patterns in the primary and secondary structure of four homologous E1a mRNAs of the adenovirus. Our main results are as follows: (1) The similarity of the coding regions of the mRNA sequences reflects both similarity in function (i.e., oncogenicity) and evolutionary divergence. (2) The similarity of the leader and the trailer regions reflects host specificity (i.e., human or simian) and must therefore arise from convergence. (3) Minimal energy foldings of the mRNAs show similar secondary structures (in particular around the splice sites). The conservation of pre-mRNA secondary structure shows that mRNAs are subject to selection constraints in addition to those associated with proteins. (4) The conserved secondary (helical) structures consist of nonhomologous subsequences, i.e., shifts have occurred. The observed shifts near the splice sites seem to be the simplest way of dealing with the dual constraints.

Introduction

Informatic molecules in biotic systems are subject to many constraints. In this paper we investigate whether, in addition to coding, there are secondary structural constraints on (pre-)mRNAs. Selection pressures can be detected by comparing the similarities between secondary structures with those between the corresponding primary structures. Sequences that have nearly identical primary structures may nevertheless have very different secondary structures, and vice versa. If minimal energy foldings of RNAs are similar, one might expect that the secondary structure is important for the functioning of the RNA and is maintained by selection.

We examined these structural features on four homologous E1a mRNA sequences of the adenoviruses: Ad5, Ad7, Ad12 (human), and SA7P (simian [African green monkey, Cercopithecus aethiops]). The expression of the E1a transcription unit, together with that of E1b, is required for complete transformation and oncogenicity of the host cell (reviewed by Van der Eb and Bernards [1984]). We show that their evolutionary relationship, as seen in their primary structure, reflects function rather than host.

The minimal energy foldings of these pre-mRNAs show that they have certain secondary structures in common. These similarities suggest that there are multilevel selectional constraints on the secondary structure of the mRNA.

Material and Methods

We have chosen the E1a mRNA of four adenoviruses as the subject of our study primarily because their relatively small size (~1,000 nucleotides) makes minimal

1. Key words: adenovirus E1a, mRNA minimal energy folding, mRNA evolutionary constraints, sequence alignment, adenovirus phylogeny, homology assessment.

Address for correspondence and reprints: Dr. P. Hogeweg, Bioinformatica, Padualaan 8, De Uitof, Utrecht, The Netherlands.

© 1987 by The University of Chicago. All rights reserved.
0737-4038/87/0403-0008$02.00
300
energy foldings of the entire mRNA feasible. In addition, conflicting "functional" patterns can be expected in the sequences because (1) three of them (Ad5, Ad7, and Ad12) are human adenoviruses and one (SA7P) is a simian (African green monkey [Cercopithecus aethiops]) adenovirus and (2) the degree of oncogenicity varies: Ad12 and SA7P are highly oncogenic; Ad7 is weakly oncogenic; and Ad5 is nononcogenic (oncogenicity is partly related to the Ela mRNA; Van der Eb and Bernards 1984).

The four adenovirus DNA sequences coding the Ela's have been determined: Ad5 (Perricaudet et al. 1979; Van Ormondt et al. 1980a; Bos et al. 1981; it is almost identical to Ad2), Ad7 (Dijkema et al. 1982), Ad12 (Perricaudet et al. 1980; Sugisaki et al. 1980), and SA7P (Dekker et al. 1984).

All four Ela pre-mRNAs can be processed to yield two main splice products—13S and 12S mRNA—early in infection. In addition, for the human adenoviruses, a 9S mRNA is known that mainly appears late in infection (for SA7P, no data are available concerning this latter splice product) (Spector et al. 1978; Chow et al. 1979; Ziff 1980; Wilson and Darnell 1981; D. Kimmelman, personal communication).

We have analyzed the pre-mRNA sequences from the 5' cap site down to the 3' poly-A attachment site, since it is known that the processing of the read-through mRNA by polyadenylation is usually the first event after transcription (Nevins and Darnell 1978). Thus, the sequence considered is the one relevant to splicing and, after splicing, to translation.

Integrated Tree-Construction, Sequence-Alignment Procedure

Sankoff et al. (1972) were the first to use an estimated genealogical relationship among sequences to assist in aligning multiple sequences. This is appropriate, since an optimal alignment is biologically ill defined without reference to the tree of their relationship. A modified Sankoff et al. procedure was outlined earlier (Hogeweg and Hesper 1984b). That procedure, called TRIALS, modified and improved still further, is detailed here. Overall, it involves estimating the relationships of the sequences (tree construction), aligning the sequences in the light of the estimated tree, and checking to see whether the alignment suggests a different tree. If so, reiterate; if not, the process has converged. The method has five steps:

**Step 1.**—Calculate similarities between all pairs of sequences by the Needleman and Wunsch algorithm (1970), modified à la Fitch and Smith (1983), to (1) permit gap penalties of the form $p = g + nr$ (where $p =$ the total penalty; $g =$ a penalty for their being a gap; $r =$ the penalty per residue in the gap; and $n =$ the number of residues spanned by the gap) and (2) permit penalizing gaps at the end of the sequence.

**Step 2.**—Formulate a starting tree from these similarities. We use the unweighted pair-group method of analysis of Sneath and Sokal (1973).

**Step 3.**—Realign the sequences according to the structure of the tree as follows: Each ancestral node has two immediate descendants, called sister sequences in analogy to sister taxa in taxonomy. Sister sequences are successively aligned and nodal sequences are computed so that an alignment of the set of sequences is obtained. The sister sequences are aligned by the Needleman and Wunsch (maximal match) algorithm with gap penalties of the form described in step 1. Among alignments with the same maximum similarity score, that alignment is used in which the gaps are arbitrarily located as far to the right (3') as possible. Gaps are treated as a fifth nucleotide when they already occur in the sequences to be aligned (which may be the case when two nodal sequences are aligned). However, newly formed gaps are fully penalized even
if they are introduced opposite gaps already present in the other sequences (otherwise, a profusion of gaps may arise in the later steps of the algorithm). A match was scored at full value irrespective of the ambiguity at that position in the two sequences; that is, in accordance with the idea of minimum mutation along the branches of the tree, no fractional matching was used. Nodal sequences must be generated by an algorithm equivalent to the first pass of Fitch’s parsimony procedure (1971), in which ambiguities are resolved as soon as a solution compatible with the parsimony paradigm is available.

**Step 4.**—Represent the tree with branch lengths corresponding to the number of nucleotide substitutions between nodes.

**Step 5.**—Reiterate the procedure, calculating a new set of similarities on the basis of sequence alignment, if necessary.

This alignment algorithm suffers from the same shortcoming as does any agglomerative clustering procedure: (arbitrary) alignment choices made earlier (i.e., near the tips [leaves] of the tree) are not changed in view of branches added later (Hogeweg and Hesper 1984b).

**Parameter Choices**

The gap penalty used in this study set \( r \) equal to zero so that the penalty was independent of length. The value of \( g \) was varied. End gaps were not penalized. Reiteration (step 5) was not performed.

These choices are now the default values in our algorithm. We believe that they are best for sequences having 50%–80% identity, such as the mRNAs studied here.

**Subsequences**

The mRNAs were divided into eight regions (see fig. 1) that were distinguished on the basis of coding capacity and degree of similarity. These regions were analyzed separately for the following reasons: (1) to prevent an outstanding local dissimilarity (e.g., a large gap for which it is hard to choose an appropriate weighting) from determining the initial tree and unduly influencing the final alignment, (2) to test whether the similarity differed over the various mRNA regions, and (3), if they should differ, to relate this to the conflicting functional patterns, the coding capacity, etc.

The final alignment was used to interpret the secondary structure. It is therefore essential that the alignment be derived by a uniform procedure and be independent of secondary-structure considerations.

**Secondary-structure Generation**

The generation of secondary structures by minimal energy folding is hampered by the absence of reliable free-energy estimates. The thermodynamic energy values reported in the literature vary considerably (e.g., Salser [1977] vs. D. H. Turner [personal communication]; Freier et al. 1985). Moreover, energy estimates from the correct foldings of tRNAs (Papanicolaou et al. 1984) yield yet other values. Therefore, many researchers resort to consensus folding without consideration of energy minimization.

Because no consensus in secondary structure can a priori be expected in mRNAs (unlike the case for small nuclear RNAs), we use minimal energy folding, but the occurrence of similar (consensus) secondary structures among the different sequences was used to reject alternative foldings.

The base content of the pre-mRNA sequences under consideration differs from that of most mRNAs in that the G and U bases are overrepresented, whereas most
Table 1

Base-Pair Stacking Energies

<table>
<thead>
<tr>
<th>EXTERIOR CLOSING PAIR</th>
<th>GU</th>
<th>UG</th>
<th>AU</th>
<th>UA</th>
<th>CG</th>
<th>GC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GU ........</td>
<td>-0.3</td>
<td>-0.3</td>
<td>-2.1</td>
<td>-2.1</td>
<td>-3.0</td>
<td>-4.8</td>
</tr>
<tr>
<td>UG ........</td>
<td>-0.3</td>
<td>-0.3</td>
<td>-1.8</td>
<td>-1.2</td>
<td>-2.1</td>
<td>-2.1</td>
</tr>
</tbody>
</table>

Note.—Data are the stacking energies of the interior G-U closing pairs that were used for the generation of the presented foldings. The other base-pair energies used (both the remaining stacking energies and the bulge-, hairpin-, and interior loop–destabilizing energies) are those reported by Salser (1977). Data are in kcal/mol.

mRNA sequences show an overrepresentation of G and C bases vis-à-vis A and U bases, and of C vis-à-vis G bases. Therefore, potential G-U bonds play a more important role and the generated foldings are sensitive to the G-U energy values, which are quite varied in the literature. Only when the Salser energy values (1977) were used, with the increase of G-U stacking energy shown in table 1, was a consensus secondary structure generated in the molecules. Other energy estimates—the default values of Salser (1977), Papanicolaou (1984), and others—were tested, and each generated a different folding and showed no similarity among the pre-mRNA molecules. The minimal energy folding of the E1a pre-mRNA of human adenovirus 40 (A. V. Loon, personal communication) conforms perfectly to the pattern described in this paper (data not shown).

Minimal energy foldings were generated by ENFOLD, a generalization of the maximum-match algorithm of Nussinov and Jacobson (1980) in which the nonmonotonicity occurring during minimal energy foldings is resolved via a backtracking mechanism: destabilizing loops are entered provisionally if no stabilizing bonds are available and are maintained if they lead to stabilizing helices; otherwise, they are ignored in later steps of the algorithm (Hogeweg and Hesper 1984a).

The algorithm generates not only the minimal energy foldings of the entire sequence but also those of any selected subsequence. We generated the foldings of the mRNA in steps of 50 nucleotides because the messengers are formed sequentially during the transcription process and because folding commences as soon as parts of the molecules are formed. The molecule may well be caught in a local energy minimum instead of switching to the global one, since the latter can only be attained by disruption of previously formed helices (the sequential folding may perhaps also be maintained by interaction with other molecules, e.g., interaction with snRNPs during transcription). Therefore sequential folding can be considered in addition to the global folding when one is searching for similar secondary structures based on energy minimization (see fig. 4, Ad7).

The minimal energy foldings were represented according to the conventions proposed by Hogeweg and Hesper (1984a). To minimize unwarranted suggestions regarding spatial relations and to facilitate the comparison of foldings of sequences and/or subsequences, we preserved the linear structure of the RNA molecule in the representation. An important aspect of this representation is that this simple visualization generates new characterizations of locations in the secondary structure. For example, the “highest top” represents the location of the largest hairpin-like structure, however complicated its internal structure may be.
FIG. 1.—Fig. 1A: The consensus tree and the associated alignment of the four E1a sequences of the Ad5, Ad7, Ad12, and SA7P as generated by our computer program, TRIALS. The Ad sequences are from humans, and the SA sequence is from the African green monkey (Cercopithecus aethiops). The alignment
Results

Alignment

The entire mRNA sequences did not reveal clearly distinguishable pairwise similarities (the identities between pairs of the Ad7, Ad12, and SA7P sequences were all ~58% and the identities between these three sequences and the Ad5 sequence were all ~51%). Thus, while Ad5 is more distinct, no decision could be made on that basis regarding the phylogeny of the other sequences. Therefore we investigated the coding and noncoding sequences separately.

Conservation of the reading frame requires insertions and deletions to be in multiples of three bases. Hence, the generated alignments and the associated tree(s) for the coding sequences were evaluated in terms of whether the gaps introduced were in multiples of three nucleotides (triplets). Since this is not a requirement of the algorithm, the introduction only of triplets is a significant result. The tree calculated from combining the two (coding) exons (regions 2–5 and 7 in fig. 1) was the one that produced an alignment of nucleotide sequences consistent with the preservation of the reading frame. This same tree is formed by most of the coding regions considered separately, especially by the regions of average similarity (regions 2, 4, and 7 in fig. 1). The same gap penalty \((g = 3)\) preserves the reading frame over nearly all coding regions, irrespective of the degree of similarity, when this particular tree is used; when other trees are used, gaps are introduced that do not preserve the reading frame.

We preferred the alignments (in terms of number and relative position of the gaps) of the noncoding regions (leader, intron, and trailer: regions 1, 6, and 8 in fig. 1) that were produced with a lower gap penalty—i.e., in which \(g = 2\) and each produces a different tree. The tree generated by the coding sequences appeared to be the best common base for the alignment of the three noncoding regions. From this we conclude that this (consensus) tree most likely represents the phylogeny of the viruses.

The consensus tree and the associated alignment of the four E1a sequences as produced by the TRIALS algorithm are shown in figure 1. The sequences of Ad12 and SA7P are clearly the most similar. Since these are sequences of the highly oncogenic adenoviruses, the evolutionary relationship of these viruses reflects the degree of oncogenicity rather than the viral host (human vs. ape). Note, however, that the rooting of the tree is based on similarity and may not necessarily reflect time of divergence.

The above-cited alignment was produced on the basis of the regions indicated in figure 1 (regions 1–8), as well as on the basis of the entire mRNA. Because it aligned the functional pyrimidine stretches downstream of the splice acceptor, we much preferred the alignment of the noncoding intron region (region 6) generated on the basis of the (consensus) tree of the coding sequences to that generated on the basis of a tree.
calculated from the similarity in this region. The leader and trailer regions are also satisfactorily aligned on this consensus tree. Although each noncoding region generated different similarity trees when analyzed separately, both the leader and trailer regions reveal similarities reflecting host specificity: SA7P is clearly separate from the others. The (noncoding) intron region, however, does not seem to reflect this host relationship; thus, the intron seems to be tuned less to a host-specific transcription/processing system.

Since a good alignment of coding and noncoding sequences can be derived from the tree on the basis of coding sequence whereas the trees generated by the noncoding sequences produce nontriplet alignments of the coding sequences and a profusion of gaps in the noncoding sequences, we conclude that the tree based on the coding sequences reflects evolutionary divergence common to all stretches and that (other) local similarity structures are the result of convergence.

The alignment obtained is essentially the same as the one that we presented recently for the deduced amino acid sequences (see Dekker et al. 1984), but there are some small differences in the location of gaps (particularly in region 4). The alignment improves the one of Van Ormondt et al. (1980b) for the same three human adenoviruses, particularly in that the gaps are in triplets over the entire coding region (see fig. 1, region 4).

The alignment has the following interesting features:

1. The 12S splice donors of Ad5 and of SA7P are located on sites that are not homologous to that of Ad7 and Ad12. However, the alignment shows that the Ad5 and SA7P sequences possess, at the site homologous to the 12S splice donor of Ad7 and Ad12, "remains" of a consensus donor; but both these remains are less complementary to the U1-interaction site involved in the splicing process than is the actual splice donor. Note that the transposition of splice sites (particularly the 12S donor site) does not reflect the evolutionary divergence, whereas parallels in length and hypervariability of the sequences do (cf. the results of Perricaudet et al. [1980]). This result might be connected with the suggestion that the 12S mRNA protein product may not be crucial: no specific function for it has been detected (Ricciardi et al. 1981; Montell et al. 1982).

2. The large gaps, as well as the regions of hypervariability as shown in the alignment, are associated with the location of the splice sites (i.e., the 12S and 13S splice donor and the splice acceptor; see also the main splice area [discussed below and shown in fig. 2]).

3. The Ad5 sequence shows a duplication in the region downstream from the splice acceptor (region 7), the 5' element (5'D in fig. 1) of which forms the stretch aligned with the large gaps in the other three sequences: the single (unique) element of the other mRNAs aligns with a slightly better fit to the 3' element (3'D in fig. 1) of the Ad5 duplicate (i.e., in all four cases, the 3' terminal coding region). This alignment corresponds with the alignment of the deduced amino acid sequences (Dekker et al. 1984). Two small subparts of Ad7, Ad12, and SA7P (yy and zz in fig. 1) have been aligned to intermediate parts of the 5' element of the duplicate in the Ad5 sequence. This fit, (fig. 1A), is only slightly better (in terms of number of gaps and substitutions) than the preferable in-reading frame alignment flanking the 5' element of the duplicate of Ad5. This alternative is shown in figure 1B.

4. The region 3' flanking the start codon (region 2) is so heterogeneous in its nucleotide and amino acid composition that the region aligns partly out of the reading frame.
FIG. 2.—Detail of the region closed off by the main stem, including the "early" used splice area. A, Ad7; B, Ad12; C, SA7P; and D, Ad5. By including the gap stretches as single-stranded parts, the primary and the secondary structure can be represented in direct relation. Similarity of the primary sequence is higher than average (■), average (=), or nonexistent (—). Broken lines (---) indicate gaps in the alignment (see fig. 1).
Fig. 3.—Minimal energy foldings of the four homologous E1a pre-mRNAs. The foldings are represented as described by Hogeweg and Hesper (1984a). Each base pair is shown by a horizontal line whose length spans the distance of that representation. Hairpin loops appear as flat tops, interior loops and bulges as intermediate plateaux, helices as sloping hillsides, and branching regions as valleys. The sequences and their calculated free energies are as follows: fig. 3A: Ad7, -384.1 kcal (the subsequences 1–250 and 251–999 were folded separately; see Material and Methods); fig. 3B: Ad12, -369.7 kcal; fig. 3C: SA7P, -454.7 kcal; and fig. 3D: Ad5, -496.3 kcal. L = Leader hairpin; ED = "early" donor hairpin; A = acceptor hairpin; ↓ = location of initiation AUG and of splice sites (13S and 12S = "early" splice donors; 9S = "late" splice donor; and SA = splice acceptor, [see Material and Methods]). Vertical lines within the pattern show the point midway between the nucleotides paired by the horizontal line.
Secondary Structure of the mRNA

The Ad5, Ad12, and SA7P secondary structures shown in figure 3 are the foldings corresponding to the overall energy minimization. The free energies are \(-496.3\) kcal for Ad5, \(-369.7\) kcal for Ad12, and \(-454.7\) kcal for SA7P. The minimal energy folding of the entire Ad7 sequence was dissimilar to the other foldings in that, apart from the leader, it revealed only one main substructure (hill); this was not subdivided in the same way as it was in the other sequences. On studying the sequential folding (discussed in the following section), we found that this divergence did not occur until the last step. Prior to this step, the pattern corresponds to the pattern of the other viruses. The secondary structure of Ad7, shown in figure 3, was obtained by folding the subsequences 1–250 and 251–999 separately. The total energy of these foldings is only a little less negative than the total energy of the minimal energy folding, i.e., \(-384.1\) kcal versus \(-398.4\) kcal.

The overall structures show some striking similarities, some occurring among all four structures and others between pairs of them. All four foldings show the presence of a large main substructure (the largest fold), the accessible position of the initiating AUG codon at the base of a local hairpin, and the configuration associated with the main (early-used) splice area (including the 12S and 13S donor and their common acceptor site).

The mRNA of Ad5 is \(\sim 130\) nucleotides longer than the mRNA of the other adenoviruses. According to our alignment (fig. 1), this is the result not of the presence

---

**Fig. 4.**—Sequential folding pattern of Ad12 in steps of 50 nucleotides. --- = Range of local hairpin in which \(\phi\) represents the hairpin loop. Blank stretches are either single stranded or base paired to remote sites (long-range interactions), --- = Long-range interactions in the entire pre-mRNA. Long-range interactions are less constant than local interactions in the sequential folding. The splice sites are indicated (12S and 13S = “early” splice donors; 9S = “late” splice donor; and SA = splice acceptor). Numbers above the horizontal line indicate the nucleotides of the pre-mRNA sequence with respect to the 5’ cap site.
of additional sequence at the 3' terminus but rather of an insertion just downstream from the splice acceptor. The folding shows that, after forming the common main helix, Ad5 has 100 nucleotides left; these form a separate fold and do not disturb the previously formed correspondence.

**Sequential folding.**—The sequential folding of Ad12 is shown in figure 4 as an example. It shows that local helices are maintained and that the main stem is formed stepwise after the splice area has been included. Since the structures are formed gradually and very little reshuffling occurs, we conclude that this sequential folding represents an energetically feasible path for the in vivo process. It is interesting to see that the “stable” structures surrounding the splice sites (the major A and ED hairpins [see below]) are much larger—i.e., include more nucleotides—than the average close-range interactions described by Nussinov and Tinoco (1981).

**The structure around the splice sites.**—The large folds (see fig. 3 and, for detail, fig. 2) are strikingly similar: they contain two major hairpins (ED and A hairpins shown as the two highest hills) in which the 12S and 13S splice donors and the splice acceptor are located. This splice area forms a separate substructure; it is closed off by part of the main stem (helix). The splice acceptor is located in the second hairpin (A hairpin) of the large substructure (fold) common to all sequences (see fig. 3). The local structure is much the same in all cases (see fig. 2). This invariance in the secondary structure contrasts strongly with the relatively large variation, around the splice acceptor (fig. 1), in the primary structure. The pyrimidine-rich part flanking the splice acceptor site is the most conserved part of this region (it is the consensus sequence of the nuclear mRNA acceptor). In the secondary structure this part forms the top of the A hairpin. The 12S and 13S donors are separated by the first major hairpin (ED) loop. Both donors are located in an accessible position, either in a single-stranded area or in a branching area (valley) at the base of the fold concerned. Note that only Ad7 and Ad12 share a homologous 12S splice donor location, as can be seen in the alignment (fig. 1). The 12S donor of SA7P is located even more upstream than that of Ad5. We predicted this location on the basis of the primary structure (Dekker et al. 1984), which has now been experimentally verified by D. Kimmelman (personal communication). Its location in the minimal energy folding fits the patterns observed here, notwithstanding its 5' deviation.

Minimal energy folding generates similar secondary structures in all four messengers. We therefore conclude that there must be selection for conservation of the secondary structure of the pre-mRNAs and, therefore, that this structure should perform some function consistent with the variation observed. The primary purpose of the unspliced messenger is to be spliced. The primary purpose of the spliced messenger is to be translated. The observed secondary structure of the leader sequence is clearly selected for its role in the translation process (see McReynolds et al. 1978; Lomedico et al. 1979; Konings et al. 1987); the other features may play a role in the splicing process.

**Comparison of Primary- and Secondary-Structure Similarity**

Minimal energy secondary structures can differ greatly as a result of slight differences in the sequence (see, e.g., Hogeweg and Konings 1985). The overall identity (50%-60% pairwise) of the primary structure is certainly not sufficient to cause the similarities that appear in the secondary structure. That the degree of sequence identity and the similarity of minimal energy foldings do not vary in parallel is shown (fig. 5) by the dissimilar folding patterns of region 5, which is 85% identical in sequence. In contrast, the folding patterns of the overall mRNAs are quite similar (see fig. 3), although the sequences are much more variable.
Throughout the mRNA, nonhomologous stretches are used to form corresponding secondary structures.—Figure 2 shows the interdependence of the similarity of primary and secondary structure for the main splice area, which includes the homologous region compared above. By including the gap regions as single-stranded parts, the primary and the secondary structures can be represented in direct relation (see fig. 2). The figure shows that the similarity of the secondary structure is conserved by shifts in the stretches that are used to form helices. Thus, in the course of evolution, mutations acceptable to the protein could be accommodated by small or larger shifts of base-pair stretches to conserve secondary structure. This seems to be an elegant solution.
for coping with the dual (multiple) constraints on the (one) sequence. The similarities between the configurations concerned (fig. 2), including the hypothesized shifts, are in agreement with the phyletic tree deduced from the sequences. The shifts and the tree emphasize the smallest divergence of Ad12 and SA7P (compare the range of base pairings of homologous stretches).

Discussion

We have examined patterns observable in the primary (coding as well as non-coding) and minimal energy secondary structure of four functionally identical mRNA nucleotide sequences and, in particular, the interdependence of these patterns.

On the basis of the similarity between the obtained minimal energy foldings and the protein structure, we concluded that there must be multiple constraints on mRNA sequences. No quantitative criterion is available for measuring secondary-structure similarity. However, the significance of the similarity can be further substantiated by experimenting with more or less scrambled sequences. Hybrid sequences (i.e., sequences into which the gaps and inserts from one sequence are inserted into another) and mutated sequences (in which the gaps and inserts are retained but randomly selected nucleotides are changed) all generated minimal energy foldings in which the pattern that was common to the original set of sequences was lost (mutation frequencies as low as that of the most conserved region [i.e., 15%] were tested).

Although the conservation of these structures suggest that they are under selection, no definite conclusion can be drawn relating particular secondary structures either to the different processes in which the molecule is involved or to their intramolecular interactions with associated molecules such as proteins and small-nuclear-RNP complexes. Nonetheless, the correspondence between known functional sites and the inferred structures suggests processes—such as translation, splicing, and survival (stability)—in which the structures could be involved. Furthermore, these patterns were obtained by considering the molecules in isolation.

Very heterologous regions of the nucleotide (and amino acid) sequences are associated with splice junctions. This feature, observed by others, has been related to the observation that exon junctions are generally located on the surface of the protein structure so that any variation flanking these junctions causes minimal disruption of the fundamental architecture of the molecule (Craik et al. 1982). This minimal disruption, however, is only possible if these variable flanking regions also consist of surface-exposable amino acids. In our set of sequences, the variable coding region 3' flanking the splice acceptor is used to stabilize the common configuration in the main splice region (see fig. 2). In the Ad5 mRNA, this part includes the 5' duplicate unit that we mentioned above (see Results and fig. 1). If we assume that this secondary pattern is important for the pre-mRNA, then the relevant (heterologous) stretches in the various mRNAs were primarily selected for that. A slight tuning of the corresponding amino acids has produced a hydrophilic region that causes only minor disturbance of the protein in that this region will be located on the surface.

On the other hand, the heterologous region enclosed by the 12S donors of SA7P and Ad12 (fig. 1) does not seem to play a crucial role in the secondary structure. If we relate this observation to the fact that the main (oncogenic) divergence of the adeno E1a mRNA is coupled with the first exon (5' flanking the 13S donor site), we conclude that this heterologous region could be a good candidate for a surface-exposed functional change (Craik et al. 1982, 1983). This conclusion is supported by the closest similarity of the amino acid sequence of the highly oncogenic Ad12 and SA7P (also see Dekker et al. 1984).
Both of the corresponding heterologous amino acid blocks are indeed hydrophilic and contain turn amino acids, so they are likely to be located on the surface. This feature applies over the whole acceptor-associated part that is common to all sequences and is significant for the 12S donor-associated part and for the 5' duplicate unit of Ad5. The duplicate of Ad5 has diverged so that the 5' unit is tuned to allow (1) the formation of a configuration of the pre-mRNA that is more or less the same as that of the others and (2) the surface exposure of the unit in the protein. The 3' unit, on the other hand, has evolved to be more hydrophobic than the corresponding parts of the other viruses; apparently this change was necessary to stabilize the protein.

The pattern that emerges from comparing the similarities of the primary and secondary structures of the four homologous mRNAs indicates that selectional constraints are dealt with, in so far as possible, by separate parts of the sequence. Notwithstanding the inherent globality of secondary structure (minimal energy folding), secondary structure seems to be conserved by specific stretches (which show little similarity between the sequences) that use slightly shifted parts of (highly conserved) stretches to form similar configurations in the various molecules. Since secondary-structure conservation seems to be stringent in the vicinity of splice sites, variable stretches are found near splice junctions.

Acknowledgments

We would like to thank H. Van Ormondt for helpful discussions and suggestions concerning this manuscript and S. M. McNab for suggesting language corrections.

LITERATURE CITED


WALTER M. FITCH, reviewing editor

Received May 12, 1986; revision received November 14, 1986.
Nonparametric Phylogenetic Inference from Restriction Cleavage Sites

Alan R. Templeton
Department of Biology, Washington University

Nei and Tajima (1985) recently have criticized my algorithms (Templeton 1983a, 1983b) that make phylogenetic inference from restriction-site maps. The purpose of this letter is to address their criticisms and to examine the reasons for our difference of opinion concerning the range of validity of my algorithms. These points will also be illustrated by a worked example using Hawaiian *Drosophila*.

Nei and Tajima (1985) have suggested that my discrimination between two primate phylogenies (Templeton 1983b) might be statistically flawed if the data included a double counting of some restriction sites. Overlap could arise owing to enzymes with multiple recognition sequences that include the recognition sequences of single cutters. However, in generating the scores used to test these hypotheses (table 2 of Templeton 1983b), I used only the novel informative patterns revealed by the multiple cutters. No site was counted twice. Hence, the concern raised by Nei and Tajima does not apply.

The primary criticism of Nei and Tajima concerns the range of validity of the algorithms. I have concluded (Templeton 1983b) that the algorithms are valid if the average substitution rate (\( \lambda \)) times the time of divergence (\( t \)) is \( \lambda t < 0.05 \). Nei and Tajima (1985) argue that the much smaller value of 0.01 is necessary. They justify this smaller \( \lambda t \) value on the basis of two theoretical arguments. The first concerns the difficulty, in the absence of outgroup data, of distinguishing a single convergent gain of restriction sites from a double loss. (fig. 3 of Nei and Tajima [1985] and associated text). However, the criticized algorithms were explicitly limited to data with an outgroup (Templeton 1983b). Hence, this theoretical objection lies outside the inference domain delimited by me (Templeton 1983b) and is therefore completely irrelevant to the validity of the 0.05\( \lambda t \) criterion.

The second theoretical objection raised by Nei and Tajima is that the probability of three parallel losses of a restriction site (or of one gain followed by two losses) is not necessarily smaller and can be greater than the probability of two parallel gains (or of one loss followed by a gain). These conclusions are sound. However, Nei and Tajima state: “Templeton (1983a) recognized this problem but concluded that if \( \lambda t < 0.05 \), the error introduced in a parsimony [emphasis mine] tree is small. . . . [A] lower criterion (\( \lambda t < 0.01 \)) is necessary.” However, the 0.05 criterion was specifically recommended for the algorithms of Templeton (1983b), which are not the same as the parsimony criterion used by Nei and Tajima in arriving at their conclusion (see, in particular, the discussion on pp. 227–229 and 239–240 of my article [Templeton 1983b]).

Although I used parsimony (Templeton 1983b), my use of it differed from Nei and Tajima’s use of it in the following two fundamental ways: (1) it used compatibility to restrict the application of parsimony to only a subset of the data, and (2) it used a weighting scheme that can assign different weights to patterns that involve the same number of events and the same weight to patterns that involve different numbers of
events. I will now show how these two differences invalidate the conclusion of Nei and Tajima concerning the limits of validity of my algorithms.

First, consider my algorithm for estimating the topology of a phylogeny. The estimate is based on compatibility, with the data generated by any given restriction enzyme being judged as incompatible if any multiple event has occurred. Hence, it is irrelevant whether a multiple event is regarded as two gains or three losses; both would result in the same qualitative inference of "incompatibility," and parsimony would not be invoked to make any further distinctions. Therefore, it is impossible for my phylogeny-estimation algorithm (Templeton 1983b) to be influenced by the ambiguity between double gains and triple losses. The estimation procedure has the formal mathematical property of being invariant with respect to this ambiguity.

Second, consider my test for discriminating between two hypothesized branching orders. The first step is to construct maximum-parsimony trees for each set of restriction sites cut by a particular enzyme under each of the two alternative branching orders. Although these restriction-enzyme trees are based on maximum parsimony, the actual test is based solely on a score assigned to the trees that assigns different weights to different types of events. A tree is assigned a score of 0 if no multiple events of any sort occur. For each additional event that involves the loss of a restriction site, a score of -1 is given, and for each additional event that involves the gain of a site, -2 is given. The score of the restriction-enzyme tree is, then, simply the sum of the individual site scores. A convergent gain or a loss-gain (i.e., one additional gain) contributes a -2 to the total score. This contribution is identical to the contribution of a triple loss or of a gain, double loss (two additional losses, each with a score of -1, which sum to -2). Since the test is based exclusively on these scores, the test has the formal mathematical property of being invariant with respect to the ambiguity between triple losses and convergent gains.

The only algorithm that I give (Templeton 1983b) that is not analytically invariant to triple losses versus double gains is the test of the molecular-clock hypothesis. Accordingly, I cautioned (Templeton 1983b) that this test should only be performed on contrasts for which convergences of any sort are rare. Under the recommended conditions, the objections of Nei and Tajima are not critical.

Therefore, neither of the theoretical justifications given by Nei and Tajima (1985) for the 0.01 criterion are valid for my algorithms. Another reason for the discrepancy between my recommendations and those of Nei and Tajima stems from the underlying assumptions of our respective models. Consequently, it is necessary to examine these assumptions.

The genetic-distance approach of Nei and Li (1979) and the probabilistic foundation (Templeton 1983a) of my algorithm (Templeton 1983b) both assume that substitutions at all nucleotides occur independently and with identical distributions that are homogeneous in all lineages over both the molecule and time. Despite some slightly different mathematical approaches and approximations, my results (Templeton 1983a) and those of Nei and Li (1979), Li (1981), and Nei and Tajima (1985) are all in close agreement concerning the probabilities of the various events that determine restriction-site evolution.

Nei and Li (1979) use the parametric form of these probabilities to generate expectations that serve as the basis of a genetic-distance measure. In contrast, I used these probability models only as guidelines in developing nonparametric ranking criteria (Templeton 1983a, 1983b). To achieve robustness to the underlying assumptions, I based my ranking criteria only on events whose parametric probabilities differed consistently by at least an order of magnitude in the relevant range of $\lambda t$ values ($\lambda t \leq 0.05$). In using this same criterion, the probability models of Nei and Tajima (1985) implied that convergent gains, loss-gains, triple losses, and gain–double losses should all be treated as an equivalence class in terms of the ranking procedure. As has already
been demonstrated, my ranking procedure (Templeton 1983b) does indeed treat these different events as equivalent.

My nonparametric approach (Templeton 1983b) was explicitly motivated by doubts concerning the validity of the homogeneity assumptions. The disagreement between me and Nei and Tajima over the 0.05 criterion stems in part from the latter's different views concerning the homogeneity assumption. In the Nei and Tajima (1985) model, $\lambda t$ is a set parametric value; for me (Templeton 1983b), it is an empirical average, with some portion of the data being above this average and some below. As I have shown, (Templeton 1983b, pp. 227–228), the compatibility approach effectively gives much more weight to that subset of the data with the lower realized $\lambda t$ values. Hence, when the average $\lambda t$ value is $\leq 0.05$, the actual statistical inference in my algorithm is primarily drawn from a subset of the data with realized $\lambda t$ values considerably $< 0.05$. Consequently, in practice, there is less of a discrepancy between my recommendations and those of Nei and Tajima than is implied by the latter.

One can make an empirical evaluation of the 0.05 criterion by studying a group of organisms for which much is already known about their phylogenetic relationships and that have $\lambda t$ values in the 0.05 range. This is the approach taken by DeSalle (1984), who examined mtDNA restriction-site evolution in Hawaiian picture-wing Drosophila. Since there is probably no other group of living organisms for which we have as much nonmolecular data available for phylogenetic inference (Carson 1983, and references therein), Hawaiian Drosophila constitute a good empirical test case of the validity of various algorithms of phylogenetic inference.

Figure 1 shows a phylogeny for eight species of Hawaiian Drosophila, estimated by my algorithm (Templeton 1983b) on the basis of data on 23 restriction enzymes (DeSalle 1984). This phylogeny is completely consistent with the outside information available on these species. DeSalle (1984) also used my algorithm to test the ability of these data to discriminate between alternative phylogenies. With the exception of the internal branching order of the alpha lineage, the estimated branching order is statistically significant at the 5% level against alternative branching orders.

Figure 2 shows the results of applying the genetic-distance, unweighted pair-group method of analysis (UPGMA) algorithm of Nei et al. (1985) to these data. As is immediately evident from the broadly overlapping SD bars surrounding the nodes shown in figure 2, not a single node is cleanly resolved. Moreover, the UPGMA tree

![Figure 1](image-url)

**FIG. 1.**—The estimated phylogeny, as determined on the basis of my algorithm (Templeton 1983b), for eight Hawaiian Drosophila species. The nonmolecular data indicate that D. neopicta is in a different taxonomic section than the remaining species, which, in turn, are split into two lineages; the alpha lineage (D. cyrtoloma, D. melanocephala, and D. neoperkinsi) and the beta lineage (D. hemipeza, D. differens, D. planitibia, and D. silvestris). Asterisks indicate those branch orderings that, as determined on the basis of my testing algorithm (Templeton 1983b), are significant at the 5% level against alternatives.
clearly clusters *D. neopicta* within the alpha lineage despite the fact that the alpha and beta lineages are in the same section and *neopicta* is in a different section. This represents a gross phylogenetic error.

This example illustrates two additional points. First, from figure 2 it is evident that the depth of this phylogeny is \( \sim 0.06 \lambda t \) units, with only one \( \lambda t \) depth \(< 0.01\). Moreover, the \( \lambda t \) depth of the beta lineage is \( 0.03 < 0.05\). My estimation and hypothesis-testing algorithms (Templeton 1983b) gave a 100% accurate and statistically significant resolution within this range. Recall that Nei and Tajima (1985) claimed that my approach would yield erroneous inferences unless all branch lengths are \(< 0.01\) and would display lower statistical power than the genetic-distance approach. Hence, this example provides empirical support for the conclusion that my algorithms are valid and can have greater statistical power than genetic-distance approaches when \( \lambda t \) is \( \sim 0.05\).

Second, this example shows that my algorithm is more robust to deviations in the evolutionary assumption of rate homogeneity than is the genetic-distance approach, which is known to be violated in this case (DeSalle 1984). The issue of the validity of assumptions is the critical point in deciding which algorithm to use. If the homogeneity assumptions of Nei et al. (1985) hold, their algorithm represents an excellent technique for analyzing restriction-site data, even when \( \lambda t \) values are \( > 0.05\). If these assumptions are not valid, my algorithms are superior, but only if \( \lambda t \) is \( < 0.05\). Consequently, the choice of algorithms depends completely on the degree of confidence that the investigator has in the evolutionary assumptions, and this degree of confidence will undoubtedly vary from data set to data set. What is needed are more empirical examples of the sort given by DeSalle (1984), to illuminate just how frequently or infrequently the evolutionary assumptions commonly made are valid.
Acknowledgments

This work was supported by National Institutes of Health grant R01 GM31571. I would like to thank Drs. Peter Smouse and Walter Fitch for their critical reading of an earlier draft of this letter.

LITERATURE CITED


Problems Arising in Phylogenetic Inference from Restriction-Site Data¹

Masatoshi Nei* and Fumio Tajima†

*Center for Demographic and Population Genetics, The University of Texas Health Science Center at Houston; and †Department of Biology, Kyushu University, Fukuoka, Japan

One of the major problems we were concerned with (Nei and Tajima 1985) is the extent of errors that would occur in estimates of the number of restriction-site changes for the various branches of the evolutionary tree when Templeton’s (1983) method is used. Templeton (1987) addresses this problem by stating that tests of the molecular-clock hypothesis should only be performed by means of his method if “convergences of any sort are rare.” Since such tests depend on the accuracy of the estimates of restriction-site changes, we take his statement to indicate agreement with our conclusion that his method does not provide reliable estimates of the numbers of restriction-site changes unless convergences of any sort are rare.

Templeton’s (1987) letter principally concerns itself with the conditions under which his method is appropriate for testing alternative tree topologies. He concludes: “If the homogeneity assumptions of Nei et al. (1985) hold, their algorithm represents an excellent technique for analyzing restriction-site data, even when λτ values are >0.05. If these assumptions are not valid, my algorithms are superior, but only if λτ is ~ ≤0.05.” Despite the moderateness of Templeton’s conclusion, we remain unconvinced.

At the present time, the theoretical basis for testing the difference in the maximum-parsimony or compatibility score between two topologies is not well established. We have shown (Nei and Tajima 1985) that if we use Templeton’s test, his phylogeny 1 (which places the chimpanzee closer to the gorilla than to the human—and is denoted as phylogeny A in our [Nei and Tajima 1985] paper) for the human, chimpanzee, gorilla, orangutan, and gibbon is never inferior to phylogeny 4 (which represents trifurcation of the human, chimpanzee, and gorilla—and is denoted as phylogeny B in our paper) even if phylogeny 4 is the correct one. This clearly indicates that Templeton’s test is not valid for this case. This conclusion remains the same regardless of whether the ambiguity between double gains and triple losses affects Templeton’s scoring system.

Let us explain this problem in a little more detail. Let s be Templeton’s score for an enzyme for the comparison of phylogenies 1 and 4. In the Wilcoxon signed-rank test that Templeton used, s must be a random variable with a mean of zero under the null hypothesis. In other words, the maximum-parsimony estimate of mutational changes for phylogeny 1 must have the same distribution as that for phylogeny 4. Actually, this is not true, and the mean or expectation of s is positive, as we showed. Therefore, the Wilcoxon test is not applicable. In the other phylogeny comparisons that Templeton considered, s is not strictly nonnegative, but its expectation (E[s]) may not be zero under the null hypothesis. Without proof that E(s) = 0 under the null hypothesis, one cannot use the Wilcoxon test. Since s is computed from the

1. Key words: phylogenetic trees, hominoid evolution, restriction-site data.

Address for correspondence and reprints: Dr. Masatoshi Nei, Center for Demographic and Population Genetics, The University of Texas Health Science Center at Houston, P.O. Box 20334, Houston, Texas 77225.
maximum-parsimony estimate for each enzyme, \( E(s) \) is unlikely to be zero in most cases. Furthermore, it is not clear what kind of tree corresponds to the null hypothesis in each case.

When there are three species involved, it is possible to conduct a statistical test of topological differences, regarding a trifurcating tree as the null-hypothesis tree (Felsenstein 1985). Under this null hypothesis, the three species are related to one another with equal distance as long as the molecular clock applies. Rejection of this null hypothesis leads to the conclusion that a particular pair of species are more closely related to each other than to the remaining one. Li (1986) applied this method to Ferris et al.'s (1981) data for mitochondrial DNA in order to distinguish between the alternative phylogenies of the human, chimpanzee, and gorilla but could not find any significant difference between them. This indicates that Templeton’s (1983) conclusion—that is, that his phylogeny 1 is significantly better than phylogeny 2—is not warranted. Li also showed that the number of phylogenetically informative sites (31) in Ferris et al.'s data is far less than the number (~400) required to obtain, with a probability of 95%, the correct tree (see Li 1986, table 9).

When there are four or more species involved, a statistical test of topological differences becomes exceedingly difficult and there seems to be no existing method that is mathematically sound (Felsenstein 1985). (For an unrooted tree of four species, Cavender [1981] developed a nonparametric test.) This is particularly so when one drops the assumption of the molecular clock. One of the major problems for this case is the difficulty in setting up a proper null hypothesis or null-hypothesis tree.

Templeton argues that the effect of rate heterogeneity on his algorithm is unimportant. We believe that his argument should be documented by mathematical proof or computer simulation, since a verbal argument is unreliable in this type of problem.

At the present time, phylogenetic trees are often constructed using parsimony or compatibility methods. As we mentioned in our 1985 paper (p. 201), these methods seem to be useful for the case of small \( \lambda t \) values. However, parsimony or compatibility methods do not utilize data on so-called singular sites (Fitch 1977), so that it is not clear whether they are superior to distance methods. As noted in our previous paper, there are many distance methods that are applicable to the case of rate heterogeneity (see Nei 1987); and Saitou and Nei’s (1986) computer simulation indicates that some of these methods are better than parsimony or compatibility methods, at least under certain circumstances.

The final problem that Templeton discusses is an empirical test of tree-making methods. He states that the evolutionary relationship of some Hawaiian \textit{Drosophila} species has been well established by means of nonmolecular studies and that this can be used as a test case of the validity of various algorithms of phylogenetic inference. According to him, the topology of the tree constructed by his algorithm when DeSalle’s (1984) restriction-site data for mtDNA is used is in complete agreement with the known phylogeny but the topology obtained by Nei et al.’s (1985) unweighted pair-group method of analysis (UPGMA) is not.

First, we do not believe that the phylogeny of the eight Hawaiian species in question is firmly established. There are no fossil records for this group of organisms, and the evolutionary relationship that has been derived from morphological, behavioral, and chromosomal data is only hypothetical (Carson 1983).

Second, the difference in topology between the tree obtained with Templeton’s method and that obtained with ours is due to the fact that Templeton used \textit{D. neopicta} as an outgroup species for his method but not for ours. If \textit{D. neopicta} is known to be an outgroup, one should use it as such and do the UPGMA analysis for the rest of the data. The results obtained would then be the same for both methods.
Citing DeSalle (1984), Templeton states that when his algorithm is used, “with the exception of the internal branching order of the alpha lineage, the estimated branching order is statistically significant at the 5% level against alternative branching orders.” This statement seems to be incorrect, even if one accepts Templeton’s test of topological differences. DeSalle (1984) considered the eight different topologies given in figure 1 and found that tree II is significantly different only from tree V. Of course, this does not mean that each of the branching orders with asterisks in Templeton’s figure 1 (equivalent to tree II) is statistically significant against alternatives. (Strictly speaking, DeSalle’s statistical conclusion should not be applied to Templeton’s fig. 1, because the species used are not identical.) In fact, the application of Li’s (1986) three-species method to DeSalle’s data indicates that the branching orders within the β-lineage group are not statistically significant against alternatives.

It should also be noted that nonparametric tests such as Templeton’s should be less powerful than parametric tests such as Nei et al.’s (1985). Since Nei et al.’s method does not discriminate between different branching orders, it is hard to imagine how Templeton’s test can.

Finally, where polymorphism for orthologous genes persists between two speciation events, Nei (1987, p. 288–289) has shown that the correct evolutionary relationship for one allele from each of several species may not be the same as that for the species themselves. This noncongruity is more likely when the species are more closely related and when fewer orthologous loci are examined. In the case of DeSalle’s Hawaiian Drosophila, the species are closely related and only one independently evolving unit of DNA (mtDNA) was studied.

Acknowledgments

We thank R. DeSalle for sending us a copy of his Ph.D. thesis. We also thank Wen-Hsiung Li, Paul Sharp, and Clay Stephens for their comments on an earlier draft of this paper.
LITERATURE CITED

CARSON, H. L. 1983. Chromosomal sequences and interisland colonizations in Hawaiian Dro-

sci. 54:217-219.

University, St. Louis.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies with a molecular clock. Syst. Zool. 34:
152-161.

FERRIS, S. D., A. C. WILSON, and W. M. BROWN. 1981. Evolutionary tree for apes and humans

FITCH, W. M. 1977. On the problem of discovering the most parsimonious tree. Am. Nat. 111:
223-257.


NEI, M., J. C. STEPHENS, and N. SAITOU. 1985. Methods for computing the standard errors of
branching points in an evolutionary tree and their application to molecular data from humans

NEI, M., and F. TAJIMA. 1985. Evolutionary change of restriction cleavage sites and phylogenetic

SAITOU, N., and M. NEI. 1986. The number of nucleotides required to determine branching
order of three species with special reference to the human-chimpanzee-gorilla divergence. J.

TEMPLETON, A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage site
maps with particular reference to the evolution of humans and the apes. Evolution 37:221-
244.

Evol. 4:315-319.
Announcements

1987 GORDON CONFERENCE ON ORIGINS OF LIFE

Theme: “Prebiotic Chemistry to Primitive Life: Where Were the Bridges?”

A. H. Delsemme, Akron—Organic Compounds in Comet Halley
J. R. Cronin, Arizona State—Interstellar Processes and Organic Compounds in Carbonaceous Meteorites
J. Oro, Houston—Prebiotic Formation of Amphiphilic Lipids and the Role of Membranes in Chemical Evolution
C. Ponnamperuma, Maryland—Recent Experiments in the Synthesis of Prebiotic Molecules
A. L. Weber, Salk—Carbohydrates as a Source of Matter and Energy for the Origins of Life
G. E. Trantor, Oxford—Parity Violation and Biomolecular Handedness
J. Lacey, Jr., Alabama—Differential Distribution of D- and L-Amino Acids Between the 2'- and 3'-Positions of Aminoacyl Nucleotides
D. A. Usher, Cornell—Chirality and Oligonucleotide-Directed Peptide Synthesis
J. R. Knowles, Harvard—The Evolution of Enzyme Function
A. Brack, CNRS, Orleans—Polypeptide Emergence and Catalysis
L. E. Orgel, Salk—Template-Directed Synthesis
A. Rich, MIT—To be announced
R. Rigler, Karolinska Institute—Structure and Dynamics of Codon-Anticodon Interactions
M. Eigen, Max Planck, Gottingen—Concepts of Sequence Space and Quasi-Species
G. E. Fox, Houston—RNA Structure Spaces and Evolution
W. F. Gilbert, Harvard—Exon Shuffling
B. F. Lang, Montreal—Introns Are Transposable in Mitochondria
M. Belfort, New York Health Department, Albany—Bacteriophage T4 Introns
E. Stackebrandt, Kiel—Prokaryotic Phylogeny
M. Sogin, National Jewish Center, Denver—Eukaryotic Phylogeny
J. Boothroyd, Stanford—Trypanosome Introns
J. Szostok, Massachusetts General Hospital—Mutational Analysis of the Self-Splicing of Introns from Tetrahymena
H. D. Robertson, Rockefeller—Evolution and Action of Viroids and Viroid-like Agents

August 17–21, 1987
Salve Regina College, Newport, Rhode Island
Registration, room and board for the conference: $325.00. Poster session will be held. For further information, see the March 6, 1987, issue of Science or contact Professor James P. Ferris, Department of Chemistry, Rensselaer Polytechnic Institute, Troy, New York 12180-3590.
ANNUAL MEETING

American Society of Naturalists (ASN)
Society for Conservation Biology
Society for the Study of Evolution (SSE)

June 21–26, Montana State University, Bozeman

Presidential Addresses:

ASN: Wyatt W. Anderson—A Case Study in Evolutionary Biology: The Inversions of *Drosophila pseudoobscura*
SSE: Douglas J. Futuyma—Isolation and Speciation in Evolutionary Biology

Symposia

*Evolution of Developmental Systems,* organized by William Atchley
  - William Atchley—Population Genetics and Development
  - Kenneth Paigen—Gene Regulation
  - Tom Kaufmann—Developmental Regulation
  - John Lucchesi—Dosage Compensation

*Paleontological Approaches to Evolutionary Dynamics,* organized by Jennifer A. Kitchell
  - Michael A. Bell—Bridging the Gap between Population Biology and Paleobiology in Stickleback Fish
  - G. P. Lohmann—Ontogeny, Ecophenotypy, and Evolution in Planktonic Foraminifera
  - David Lazarus—Speciation in the Pelagic Realm
  - Steven M. Stanley—Evolutionary Stasis in Multilineages of Neogene Bivalves
  - Jennifer A. Kitchell—Stasis and Change: The Nature of Coevolutionary Dynamics

*Sex-Allocation Research: Implications to General Questions in Evolution,* organized by John H. Werren
  - John H. Werren—Levels of Selection and Sex Allocation
  - Jon Seger—Sex Allocation and the Evolution of Insect Sociality
  - William Stubblefield—Sex Allocation in Spatially Structured Populations
  - Nancy Moran—Sex Allocation and Parthenogenetic Life Cycles

*Variation in Male Fertility in Higher Plants,* organized by James D. Thomson and Beryl B. Simpson
  - Deborah Charlesworth—Effects on Fitness of Allocation to Attractive Structures
  - David Lloyd—Contrasts between Male and Female Strategies
  - Thomas R. Meagher—Analysis of Parentage in Natural Populations
  - Barbara A. Schaal and Kaius Helenurm—Genotypic Effects on Reproductive Success in *Lupinus*
  - Beryl B. Simpson and John L. Neff—Floral Resources and Insect Foraging: Implications for Male Function in Plants
  - Daniel J. Schoen and Steven C. Stewart—Estimation of Male Fertility in Plant Populations
James D. Thomson and L. Harder—Pollen Presentation and Delivery as Components of Male Reproductive Success in *Erythronium*

**Checkerspot Butterflies as a Model System in Population Biology,** organized by Peter F. Brussard
- Paul R. Ehrlich—Checkerspot Butterflies: An Historical Perspective
- Peter F. Brussard—Genetic Variation and Systematic Relationships
- Dennis Murphy—Microhabitat Resources and Metapopulation Structure
- Deane M. Bowers—Secondary Compounds and the Coevolution of *Euphydryas* and its Host Plants
- Michael Singer—Behavioral Aspects of *Euphydryas*—Host Plant Coevolution

**Cytoplasmic-Nuclear Interactions,** organized by Jonathan Arnold
- R. A. Lansman—Molecular Mechanisms for Mitochondrial-Nuclear Interactions
- Marjorie A. Asmussen—Cytonuclear Associations: Models and Measures of Disequilibria
- Andrew G. Clark and A. F. MacRae—Mitochondrial-Nuclear Interactions in Experimental Populations
- John C. Avise—Mitochondrial-Nuclear Interactions in Hybrid Zones
- Samuel Levings—Molecular Evolution of Plant mtDNA in Concert with the Nuclear Genome

**Young Investigator's Symposium**
- Three or four recent Ph.D.'s will be selected

**The Role of Disease in Population Regulation and Conservation,** organized by Robert M. May

**Edge Effects in Conservation,** organized by Larry Harris

**Conservation Genetics of Fishes,** organized by Fred W. Allendorf

**How We Train Conservation Biologists,** organized by David Hales

For information on housing and registration, contact Dr. Peter F. Brussard, Department of Biology, Montana State University, Bozeman, Montana 59717; phone (406) 994-4548. For information on the program and contributed papers, contact Dr. Richard G. Harrison, Department of Ecology and Systematics, Section of Ecology and Systematics, Corson Hall, Cornell University, Ithaca, New York 14853; phone (607) 255-6583.
Subscribe to

BioScience

The biologist’s monthly magazine

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

Join AIBS and start your subscription today.

Name ____________________________

Address ______________________________________

____________________________________

Charge my _______ ______ Acct. # ________________________

Signature ______________________________________

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

Please make checks payable to AIBS. Remit US dollars only. Mail to: AIBS, P.O. Box 27417, Central Station, Washington, DC 20038-7417.
Whether you’re in practice or research, staying ahead can be as easy as 1, 2, 3

1. **The Journal of Infectious Diseases***

- **JID** brings you up-to-date coverage on the wide range of research that influences medical microbiology and your work, covering clinical investigations, experimental medicine, controlled trials, and research from microbiology, immunology, epidemiology, and related disciplines.

- **JID** gives you access to research as soon as it’s available—with lead times of only 6-8 months on most articles, and 2-3 months on communications of unusual interest.

- **JID** provides exposure to the field beyond your usual focus. In addition to original research, **JID** presents concise laboratory and clinical communications, correspondence on items of unusual teaching, clinical, or scientific value, and summaries of workshops.

2. **Reviews of Infectious Diseases***

- **RID** saves you the time of tracking information from several sources. **RID**'s coverage of the most recent literature in medical microbiology includes works on infectious organisms, the immune system and its diseases, and anti-microbial agents; on therapeutic advances; and on results of clinical investigations.

- **RID** makes your exposure to the field complete with review articles, ongoing examinations of work on urgent or timely problems, thematic studies, historical articles, articles on history, and case reports.

3. **Subscribe now.**

*The Journal of Infectious Diseases* is published monthly. Editor: Martha D. Yow

*Reviews of Infectious Diseases* is published bimonthly. Editor: Edward H. Kass

**JID** and **RID** are published by The University of Chicago Press.

**Individual Rates**

| Combination JID/RID subscription: individuals | $105.00; institutions $225.00; ASM & IAMS individual members $90.00; students, interns, residents, and fellows (attach copy of ID) | $50.00. Outside USA add $22.00. |
|-----------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------|
| Subscription rates for either JID or RID (be sure to indicate which journal you are ordering): individuals | $57.00; institutions $125.00; ASM & IAMS individual members $50.00; students, interns, residents, and fellows (attach copy of ID) | $30.00. Outside USA add $12.00 for JID subscription; $10.00 for RID subscription. |
| Visa or MasterCard accepted. Send charge card information, purchase order, or payment to The University of Chicago Press, Journals Division, P.O. Box 37005, Dept. SS7SA, Chicago, IL 60637. |
Since it was first published more than 25 years ago, the Journal of Molecular Biology has been at the forefront of research in the field and has helped to establish and to enlarge the boundaries of molecular biology. The journal continues to provide up-to-date and comprehensive coverage of the entire, rapidly expanding field while increased emphasis is given to those areas of cell biology in which the molecular approach is presently achieving substantial results.

Research Areas Include
- Biological structures
- Molecular genetics
- Mechanisms of replication

Publication: Bi-weekly; Subscription: Volumes 193 – 198, 1987

£660.00 (U.K.)/$1,250.00 (Overseas)

Recent Contents

<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteriophage T7 Resolves Conformationally Branched Structures in Double-stranded DNA.</td>
<td>R.B. Inman and M. Schnos</td>
</tr>
<tr>
<td>Electron Microscopic Identification of Super-coiled Regions in Complex DNA Structures.</td>
<td>P. Argos</td>
</tr>
<tr>
<td>A Sensitive Procedure to Compare Amino Acid Sequences.</td>
<td>K.M. Sullivan and D.M.J. Lilley</td>
</tr>
<tr>
<td>Influence of Cation Size and Charge on the Extrusion of a Salt-dependent Cruciform.</td>
<td>S.M. Wernke and J.B. Lingrel</td>
</tr>
</tbody>
</table>

Coupon: JMB/87/E.124

Please tick □ Please send me a sample copy and subscription details of Journal of Molecular Biology

Name (Capitalise) .................................................................................................................................

Address ...................................................................................................................................................

........................................... ........................................... ........................................... Date ..............................

A C A D E M I C  P R E S S

24-28 OVAL ROAD, LONDON NW1 7DX • 1250 SIXTH AVENUE, SAN DIEGO, CA 92101
For ecologists, geneticists, and evolutionary biologists . . .

The American Naturalist

Relating observational and theoretical discoveries in such areas as genetics, behavior, physiology, and molecular biology to the fields of population biology and evolution, The American Naturalist helps you explore the diversity of current research. Established in 1867, The American Naturalist is one of the foremost publications in the world for research in ecology, evolution, and population biology. Presenting a unified approach to the biological sciences, the journal includes empirical and theoretical reports on such topics as:

- Ecological succession
- Evolutionary genetics
- Plant breeding systems
- Migrational behavior in animals
- Animal mating systems
- Dispersal mechanisms in plants
- Adaptive behavior

The American Naturalist: A natural choice for vital research

Editor: Marcus W. Feldman  Associate editors: Paul R. Ehrlich, Harold A. Mooney, Jonathan Roughgarden, and Peter M. Vitousek

Published monthly for the American Society of Naturalists by The University of Chicago Press

One-year rates: Individuals, $50.00; institutions, $90.00; students (with copy of ID), $35.00. Outside USA, add $9.00 for postage. Visa and MasterCard accepted. Send check, purchase order, or complete charge card information to The University of Chicago Press, Journals Division, Dept. SW7SA, P.O. Box 37005, Chicago, IL 60637.

1/87
"A convenient collection of review articles . . . useful to students as a springboard for historical research."

—Sharon Kingsland, The Johns Hopkins University

BULLETIN OF THE HISTORY OF MEDICINE

Genetic Perspectives in Biology and Medicine
Edited by Edward D. Garber

An excellent and important addition to textbook lists for courses in:
- genetics
- evolution
- history of science
- philosophy of science
- biochemistry
- molecular biology

Documenting the dramatic developments in genetics during the last 25 years, this volume presents first-hand perspectives from the initiators of modern research and discovery. This collection of essays, originally published in Perspectives in Biology and Medicine, includes contributions from such distinguished scientists as Francis Crick, Bruce Wallace, E. L. Tatum, and Curt Stern. Topics addressed include the history of genetics, human genetics, the work of fellow researchers, and the future of genetic research. Insightful overviews by the editor place the articles—dating from 1957 to 1985—into their proper historical context, making the book a vital addition to your research library.

Order Form

Yes I want the book that covers the most important developments in biology since Darwin . . . Please send me ___ copies of Genetic Perspectives in Biology and Medicine (Paper ISBN: 0-226-28216-3) at $12.00 each.*

Yes, please send me an examination copy of Genetic Perspectives in Biology and Medicine (Paper ISBN: 0-226-28216-3). I understand that I will be billed at a 10% academic discount, but that the charges will be cancelled upon notification that my campus bookstore has placed an order for ten or more copies. If I do not adopt the book for my course, I may return the examination copy for a full refund or purchase it for my personal library.

*Add sales tax: Chicago 8%; elsewhere in IL 7%. Outside the Western Hemisphere prices will be slightly higher. Add $1.50 for postage and handling.

Name __________________________
Address _________________________
City/State/Zip ____________________

Visa or MasterCard accepted. Send check, purchase order, or complete charge card information to The University of Chicago Press, Journals Division, SW7SA, P.O. Box 37005, Chicago, IL 60637.
CBE Works for You

The Council of Biology Editors, Inc. serves writers, editors, and publishers in the biological sciences through its outstanding membership services and publications, including:

- **CBE Views**, a quarterly publication, keeping you informed of the latest developments in scientific communication and publishing

- publications providing valuable assistance for editors and writers, including the **CBE Style Manual**, the standard reference in the biological sciences

- an annual meeting stressing continuing education and networking among participants

Find out more about the Council of Biology Editors today!

☐ Send me more information about CBE membership.
☐ Send me more information about CBE publications.

Return to:
Council of Biology Editors, Inc., Dept. D-86, 9650 Rockville Pike, Bethesda, Maryland 20814, (301) 530-7036

Name ____________________________________________

Organization ____________________________________

Street Address ___________________________________

City, State, Zip Code _______________________________

Country __________________________________________ Phone Number ( ) ____________________________
Revised and Expanded by Bonnie Birtwistle Honigsblum

Extensively revised and issued in a larger format, this new edition retains the features that have made the manual, in its earlier versions, the choice of over 5,000,000 writers and students. Among the new features are a focus on computer usage for the preparation of manuscripts; an updated guide to standard reference styles, with a simplified form of parenthetical reference; a bibliography of style guides recommended for various academic disciplines; expanded information on the preparation and presentation of tables; extensive new examples throughout the text and a new chapter of twenty-eight annotated sample pages.

Cloth $20.00  Paper $7.95  302 pages (est.)  9 tables, 26 line drawings

Chicago Guides to Writing, Editing, and Publishing

The University of CHICAGO Press
5801 South Ellis Avenue, Chicago, IL 60637
Forthcoming

Origin of Structural Domains of the Serum Albumin Gene Family and a Predicted Structure of the Gene for Vitamin D-binding Protein
Peter E. M. Gibbs and Achilles Dugaiczyk

A Statistical Analysis of Nucleotide Sequences of Introns and Exons in Human Genes
Michael Bulmer

Intraspecific DNA Divergence in Drosophila: A Study on Parthenogenetic D. mercatorum
Adalgisa Caccone, George D. Amato, and Jeffrey R. Powell

Codon Contexts in Enterobacterial and Coliphage Genes
Manolo Guoy

Distributions of Two Recently Inserted Long Interspersed Elements of the L1 Repetitive Family at the A1b and βh3 Loci in Wild Mice Populations
Michel Bellis, Véronique Jubier-Maurin, Barbara Dod, Flavie Vanlerberghe, Anne-Marie Laurent, Christiane Senglat, François Bonhomme, and Gérard Roizès

The Evolution of Prokaryotic Ferredoxins—with a General Method Correcting for Unobserved Substitutions in Less Branched Lineages
Walter M. Fitch and Mireille Bruschi

L1 Repeat Elements in the Human ε-γ-Globin Gene Intergenic Region: Sequence Analysis and Concerted Evolution within This Family
Peter K. Rogan, Julian Pan, and Sherman M. Weissman
Review Article

203 Slipped-Strand Mispairing: A Major Mechanism for DNA Sequence Evolution
Gene Levinson and George A. Gutman

Major Articles

222 The Rate of Synonymous Substitution in Enterobacterial Genes Is Inversely Related to Codon Usage Bias
Paul M. Sharp and Wen-Hsiung Li

231 Complete Nucleotide Sequences of Bovine αS2- and β-Casein cDNAs: Comparisons with Related Sequences in Other Species
A. Francis Stewart, John Bonsing, Craig W. Beattie, Farida Shah, Ian M. Willis, and Antony G. Mackinlay

242 DNA-DNA Hybridization Evidence of the Rapid Rate of Muroid Rodent DNA Evolution
François M. Catzeflis, Frederick H. Sheldon, Jon E. Ahlquist, and Charles G. Sibley

254 Restriction-Map Variation in the Alcohol Dehydrogenase Region of Drosophila pseudoobscura
Stephen W. Schaeffer, Charles F. Aquadro, and Wyatt W. Anderson

266 Temperature-dependent Kinetic Variation among Phosphoglucose Isomerase Allozymes from the Wing-polymorphic Water Strider, Limnoporus canaliculatus
Anthony J. Zera

287 Evolutionary Diversification of Class II P Loci in the Mhc of the Mole-Rat Spalax ehrenbergi
Ralf Schöpfer, Felipe Figueroa, Dean Nižetić, Eviatar Nevo, and Jan Klein

300 Evolution of the Primary and Secondary Structures of the E1a mRNAs of the Adenovirus
D. A. M. Konings, P. Hogeweg, and B. Hesper

Letters to the Editor

315 Nonparametric Phylogenetic Inference from Restriction Cleavage Sites
Alan R. Templeton

320 Problems Arising in Phylogenetic Inference from Restriction-Site Data
Masatoshi Nei and Fumio Tajima
Announcements

324 1987 Gordon Conference on Origins of Life

325 Annual Meeting: American Society of Naturalists, Society for Conservative Biology, and Society for the Study of Evolution

Molecular Biology and Evolution (ISSN 0737-4038) is published six times a year in January, March, May, July, September, and November by the University of Chicago Press, 5801 Ellis Avenue, Chicago, Illinois 60637, and sponsored by the Molecular Biology and Evolution Society and the American Society of Naturalists. Molecular Biology and Evolution is indexed in Current Contents/Life Sciences, Science Citation Index, and Excerpta Medica.

Subscription Rates U.S.A.: institutions, 1 year $95.00; individuals, 1 year $35.00. Student subscription rate, U.S.A.: 1 year $28.00 (copy of student ID must accompany subscription). Other countries add $4.00 for each year's subscription to cover postage. Subscription agent for Japan: Kinokuniya Company, Ltd. Single copy rates: institutions $15.85, individuals $5.85. Business correspondence should be addressed to The University of Chicago Press, Journals Division, P.O. Box 37005, Chicago, Illinois 60637.

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

Editorial Correspondence Correspondence about book reviews should be addressed to the Managing Editor: Masatoshi Nei, Center for Demographic and Population Genetics, University of Texas Health Science Center at Houston, P.O. Box 20334, Houston, Texas 77225. Editorial correspondence should be addressed to the Editor in Chief: Walter M. Fitch, Department of Biological Sciences, University of Southern California, University Park, Los Angeles, California 90089-1482. For manuscript submission, see Information for Contributors at the end of this issue.

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

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

© 1987 by The University of Chicago. All rights reserved.