Review Article

Slipped-Strand Mispairing: A Major Mechanism for DNA Sequence Evolution[1]

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 1984*a*, 1984*b*). 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. 1982*a*, 1982*b*; Rogers

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: $[GATA]_n$ and $[GACA]_n$ 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.

Mol. Biol. Evol. 4(3):203–221. 1987.

0737-4038/87/0403-0001$02.00

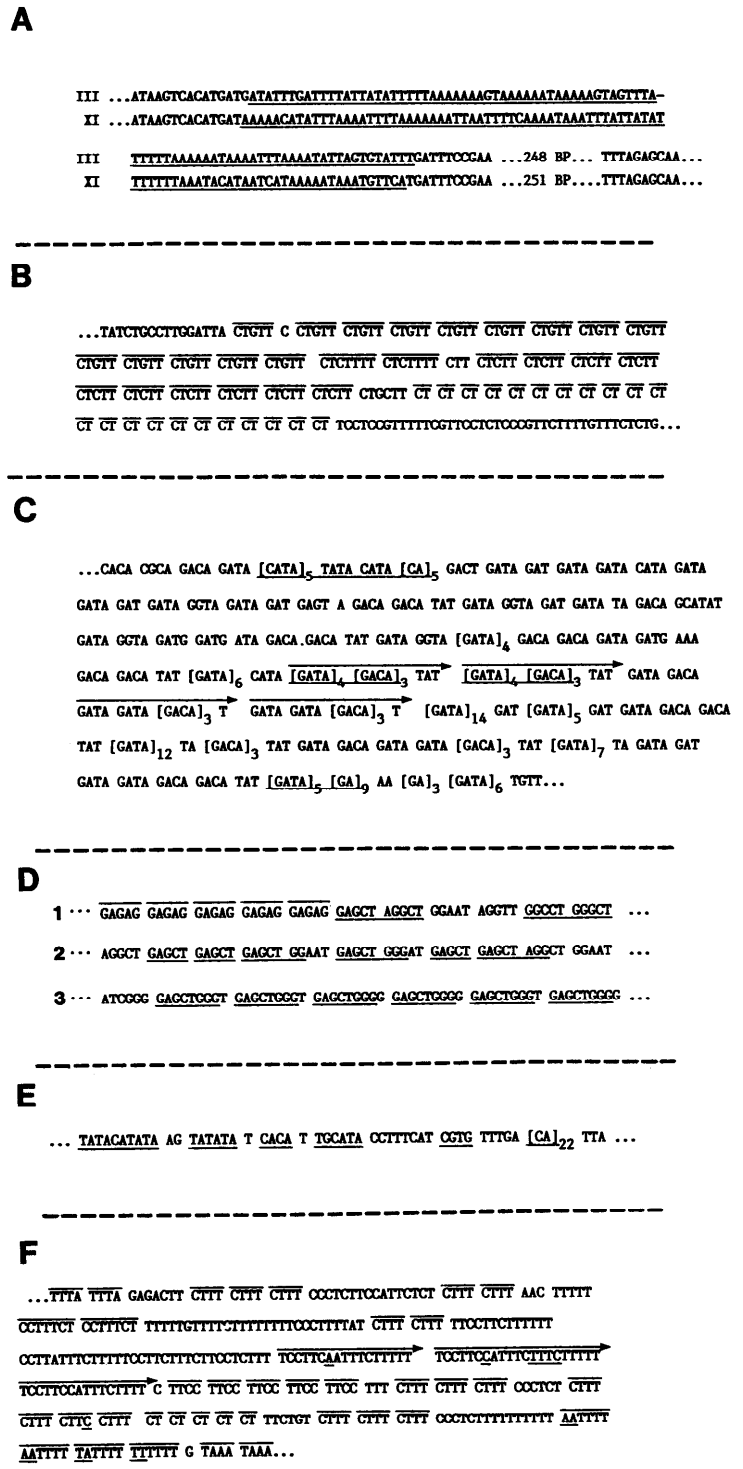

FIG. 1.—Naturally occurring simple repetitive sequences. All sequences represent duplex DNA, but only one strand is shown. Included are examples of runs of a single base (A); tandem reiterations with repeats of ≥2 bases (B, C, and D); imperfect (quasi-) reiterated sequences (D, E); and sequences with 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 Hughs 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. 1982*a*, 1982*b*; Miklos and Gill 1982; Rodakis and Kafatos 1982; Singer 1982; Sorge and Hughs 1982; Moore 1983; Cato et al. 1984; Hasson et al. 1984; Rodakis et al. 1984; Skowronski 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-repetitive 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 $GAGCTGGG^T/_G$; 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 ≥ 8 bases and with repeat units of length ≥ 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 $\geqslant 10$ 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 $[A]_4$ and $[A]_5$ 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 Hughs 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

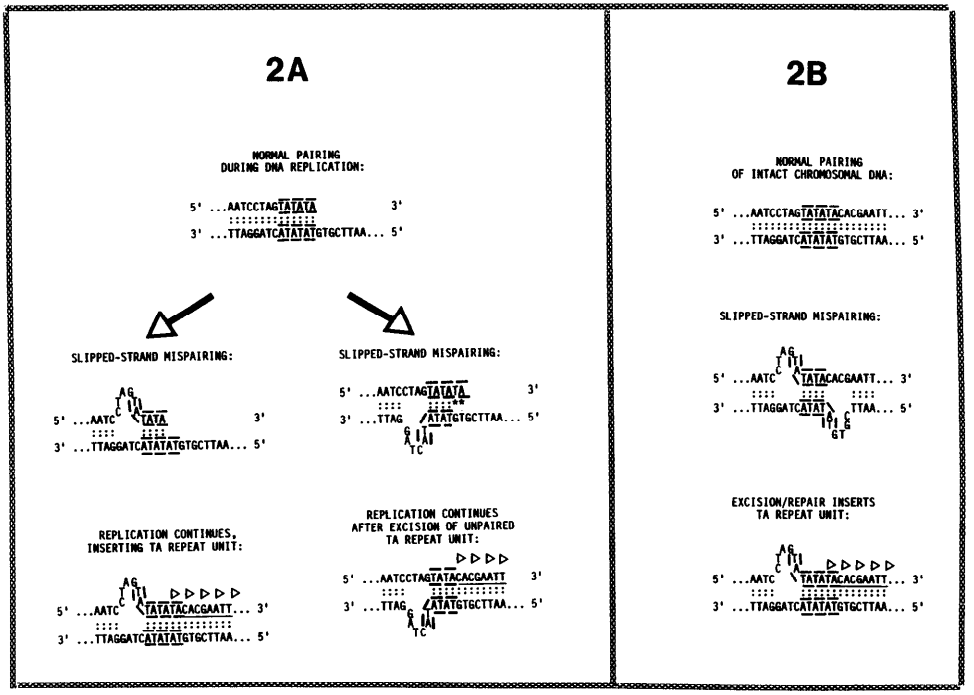

FIG. 2.—Generation of duplications or deletions by SSM between contiguous repeats. Small arrows indicate direction and starting point of DNA synthesis; colons indicate base pairing. A, 2-Base slippage in an AT-repeat during replication of a DNA duplex, followed by continued chain elongation. Slippage 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 polypyrimidine tracts, containing a variety of perfect and imperfect tandem repeats, would tend to result.

The prevalence of poly-CA tracts (Hamada et al. 1982*a*, 1982*b*; 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 polypyrimidine 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 ≥ 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 (1967*a*, 1967*b*) 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 Renz 1984*a*, 1984*b*), duplications and deletions (Efstradiatis et al. 1980) including coupled events (Flanagan et al. 1984), gene conversion (Slightom 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. (1967*b*) 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

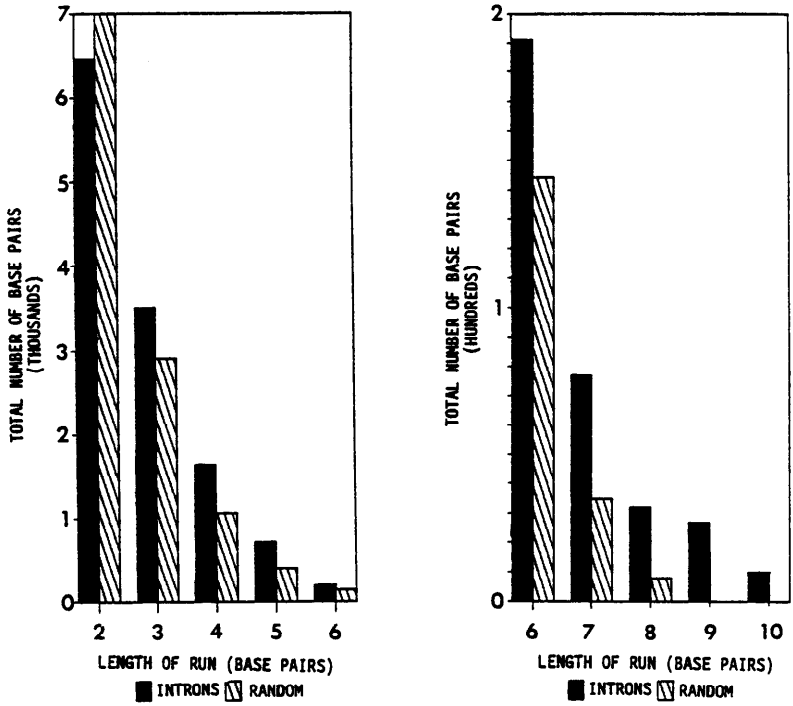

FIG. 4.—Frequency distributions of runs of a single base in 91 mammalian introns from 25 genes (see Appendix A) and their pseudorandom counterparts. The total number of bases included in runs of length 2–10 is indicated by the bars; data for introns are represented by solid bars, those for their random counterparts by cross-hatched bars. The right-hand graph has an ordinate that is expanded relative to that on the left, and the data for length category 6 is shown in both graphs.

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 (Slightom 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

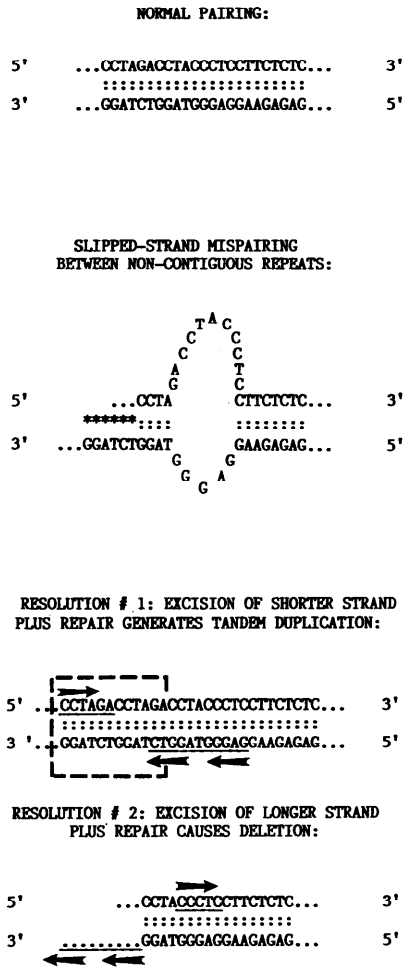

FIG. 5.—SSM involving *noncontiguous* repeats, which can generate deletions or duplications of the region included between the mispairing regions. Depending on whether the mispaired structure is resolved by excision of the longer or shorter single-stranded loop, the result can be duplication (box, Resolution #1) or deletion (Resolution #2) of the region spanned by the mispaired CCTA sequences. Underlined bases are replaced by the excision/repair system during resolution.

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 S_1 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. 1984*a*). 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 eukarytic 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. (1984*b*) 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. (1984*b*) 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. 1984*a*) 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; BOVOPS1-5, bovine opsin; BOVPOMC3-5,7, bovine proopiomelanocortin; DOGINS, dog insulin; GOTHBAI, goat adult alpha-i-globin; HAMVIM1-7, hamster vimentin; HUMA1AT1-4, human alpha-1 antitrypsin; HUMACTCA1-4, human alpha-cardiac actin; HUMAPOAI1, 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-amylase-1; MUSFOL1-6, mouse dihydrofolate reductase; MUSIGDJC10, mouse immunoglobulin germ-line d-j-c region: mu; RABHBB1A1, rabbit beta-1 globin; RATAVP1-2, rat arginine vasopressin-neurophysin precursor; RATCASG11-12, 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 ≥ 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

ANDERSON, R. P., and J. ROTH. 1977. Tandem genetic duplications in phage and bacteria. Annu. Rev. Microbiol. **31**:473–505.

———. 1981. Spontaneous tandem genetic duplications in *Salmonella typhimurium* arise by unequal recombination between rRNA (rrn) cistrons. Proc. Natl. Acad. Sci. USA **78**:3113–3117.

APPELS, R., and W. J. PEACOCK. 1978. The arrangement and evolution of highly repeated (satellite) DNA sequences with special reference to *Drosophila*. Int. Rev. Cytol., Suppl. 8: 69–126.

BLAISDELL, B. E. 1983. A prevalent persistent global nonrandomness that distinguishes coding and non-coding eucaryotic nuclear DNA sequences. J. Mol. Evol. **19**:122–133.

BLOOM, K. S., M. FITZGERALD-HAYES, and J. CARBON. 1982. Structural analysis and sequence organization of yeast centromeres. Cold Spring Harbor Symp. Quant. Biol. **46**:1175–1185.

BROWN, W. M., E. M. PRAGER, A. WANG, and A. C. WILSON. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. **18**:225–239.

BRUTLAG, D. L. 1980. Molecular arrangement and evolution of heterochromatic DNA. Annu. Rev. Genet. **14**:121–144.

CARLSON, M., and D. BRUTLAG. 1979. Different regions of a complex satellite DNA vary in size and sequence of the repeating unit. J. Mol. Biol. **135**:483–500.

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.

CHAPMAN, B. S., K. A. VINCENT, and A. C. WILSON. 1986. Persistence or rapid generation of DNA length polymorphism at the zeta-globin locus of humans. Genetics **112**:79–92.

COULONDRE, C., J. H. MILLER, P. J. FARABAUGH, and W. GILBERT. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. Nature **274**:775–780.

DAVIS, M. M., S. K. KIM, and L. E. HOOD. 1980. DNA sequences mediating class switching in alpha-immunoglobulins. Science **209**:1360–1365.

DEBOER, J. G., and L. S. RIPLEY. 1984. Demonstration of the production of frameshift and base-substitution mutations by quasipalindromic DNA sequences. Proc. Natl. Acad. Sci. USA **81**:5528–5531.

DODD, J. G., and N. A. STRAUS. 1982. Repeated sequences in L-cell mRNA complementary to long deoxypolypyrimidines. Biochim. Biophys. Acta **698**:140–148.

DOOLITTLE, W. F., and C. SAPIENZA. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature **284**:601–603.

EFSTRADIATIS, A., J. W. POSAKONY, T. MANIATIS, R. M. LAWN, C. O'CONNELL, R. A. SPRITZ, J. K. DERIEL, B. G. FORGET, S. W. WEISSMAN, J. L. SLIGHTOM, A. E. BLECHL, O. SMITHIES, F. E. BARALLE, C. C. SHOULDERS, and N. J. PROUDFOOT. 1980. The structure and evolution of the human β-globin gene family. Cell **21**:653–668.

EPPLEN, J. T., A. CELLINI, S. ROMERO, and S. OHNO. 1983. An attempt to approach the molecular mechanisms of primary sex determination: w- and y-chromosomal conserved simple repetitive DNA sequences and their differential expression in mRNA. J. Exp. Zool. **228**:305–312.

FARABAUGH, P., U. SCHMEISSNER, M. HOFER, and J. H. MILLER. 1978. Genetic studies of the lac repressor. VII. On the molecular nature of spontaneous hotspots in the *lac I* gene of *Escherichia coli*. J. Mol. Biol. **126**:847–863.

FLANAGAN, J. G., M.-P. LEFRANC, and T. H. RABBITTS. 1984. Mechanisms of divergence and convergence of the human immunoglobulin α-1 and α-2 constant region gene sequences. Cell **36**:681–688.

FOWLER, R. G., G. E. DEGNEN, and E. C. COX. 1974. Mutational specificity of a conditional *Escherichia coli* mutator, mutD5. Mol. Gen. Genet. **133**:179–191.

FRESCO, J. R., and B. M. ALBERTS. 1960. The accommodation of noncomplementary bases in helical polyribonucleotides and deoxyribonucleic acids. Proc. Natl. Acad. Sci. USA **46**:311–321.

GILLIES, S. D., V. FOLSOM, and S. TONEGAWA. 1984. Cell type-specific enhancer element associated with a mouse MHC gene, E_β. Nature **310**:594–597.

GLICKMAN, B. W. 1981. Methylation-instructed mismatch correction as a postreplication error avoidance mechanism in *Escherichia coli*. Pp. 65–87 *in* J. F. LEMONTT and W. M. GENEROSO, eds. Molecular and cellular mechanisms of mutagenesis. Plenum, New York.

GROSSMAN, L. 1981. Enzymes involved in the repair of damaged DNA. Arch. Biochem. Biophys. **211**:511–522.

HAMADA, H., and T. KAKUNAGA. 1982*a*. Potential Z-DNA sequences are highly dispersed in the human genome. Nature **298**:396–398.

HAMADA, H., M. G. PETRINO, and T. KAKUNAGA. 1982*b*. A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. Proc. Natl. Acad. Sci. USA **79**:6465–6469.

HAMADA, H., M. G. PETRINO, T. KAKUNAGA, M. SEIDMAN, and B. D. STOLLAR. 1984*a*. Characterization of genomic poly(dT-dG)·poly(dC-dA) sequences: structure, organization, and conformation. Mol. Cell. Biol. **4**:2610–2621.

HAMADA, H., M. SEIDMAN, B. H. HOWARD, and C. M. GORMAN. 1984*b*. Enhanced gene expression by the poly(dT-dG)·poly(dC-dA) sequence. Mol. Cell. Biol. **4**:2622–2630.

HAMORI, E. 1975. Kinetic investigation of unfolding and partial refolding of a crab satellite $(dA\text{-}dT)_n$. Biochemistry **14**:3618–3625.

HANAWALT, P. C., P. K. COOPER, A. K. GANESAN, and C. A. SMITH. 1979. DNA repair in bacteria and mammalian cells. Annu. Rev. Biochem. **48**:783–836.

HASSON, J.-F., E. MOUGNEAU, F. CUZIN, and M. YANIV. 1984. Simian virus 40 illegitimate recombination occurs near short direct repeats. J. Mol. Biol. **177**:53–68.

HEILIG, R., R. MURASKOWSKY, and J.-L. MANDEL. 1982. The ovalbumin gene family: the 5′ end region of the X and Y genes. J. Mol. Biol. **156**:1–19.

HENTSCHEL, C. C. 1982. Homocopolymer sequences in the spacer of a sea urchin histone gene repeat are sensitive to S_1 nuclease. Nature **295**:714–716.

HOLMQUIST, R. 1983. Transitions and transversions in evolutionary descent: an approach to understanding. J. Mol. Evol. **19**:134–144.

HSIEH, T.-S., and D. BRUTLAG. 1979. Sequence and sequence variation within the 1.688 g/cm^3 satellite DNA of *Drosophila melanogaster*. J. Mol. Biol. **135**:465–481.

INGLE, J., G. G. PEARSON, and J. SINCLAIR. 1973. Species distribution and properties of nuclear satellite DNA in higher plants. Nature New Biol. **242**:193–197.

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

of tandemly repeated CA dinucleotides hybridizes to highly repetitive dispersed elements in mammalian cell genomes. Mol. Cell. Biol. **3**:1389–1402.

JEFFREYS, A. J., V. WILSON, and S. L. THEIN. 1985. Individual-specific 'fingerprints' of human DNA. Nature **316**:76–79.

JONES, C. W., and F. C. KAFATOS. 1982. Accepted mutations in a gene family: evolutionary diversification of duplicated DNA. J. Mol. Evol. **19**:87–103.

KORNBERG, A. 1980. *In* DNA replication. W. H. Freeman, San Francisco.

KORNBERG, A., L. L. BERTSCH, J. F. JACKSON, and H. G. KHORANA. 1964. Enzymatic synthesis of deoxyribonucleic acid. XVI. Oligonucleotides as templates and the mechanisms of their replication. Proc. Natl. Acad. Sci. USA **51**:315–323.

KRAMER, B., W. KRAMER, and H.-J. FRITZ. 1984. Different base/base mismatches are corrected with different efficiencies by the methyl-directed DNA mismatch-repair system of *E. coli*. Cell **38**:879–887.

KRAMER, W., K. SCHUGHART, and H.-J. FRITZ. 1982. Directed mutagenesis of DNA cloned in filamentous phage: influence of hemimethylated GATC sites on marker recovery from restriction fragments. Nucleic Acids Res. **10**:6475–6485.

KRONENBERG, M., G. SIU, L. E. HOOD, and N. SHASTRI. 1986. The molecular genetics of the T-cell antigen receptor and T-cell antigen recognition. Annu. Rev. Immunol. **4**:529–591.

KUNKEL, T. A. 1985. The mutational specificity of DNA polymerase-beta during *in vitro* DNA synthesis: production of frameshift, base substitution, and deletion mutations. J. Biol. Chem. **260**:5787–5796.

KURNIT, D. M. 1979. Satellite DNA and heterochromatin variants: the case for unequal mitotic crossing over. Hum. Genet. **47**:169–186.

LEE, C. S., R. W. DAVIS, and N. DAVIDSON. 1970. A physical study by electron microscopy of the terminally repetitious, circularly permuted DNA from the coliphage particles of *Escherichia coli* 15. J. Mol. Biol. **48**:1–22.

LEVIN, D. E., E. YAMASAKI, and B. N. AMES. 1982. A new Salmonella tester strain, TA97, for the detection of frameshift mutagens: a run of cytosines as a mutational hot-spot. Mutat. Res. **94**:315–330.

LEVINSON, G., J. L. MARSH, J. T. EPPLEN, and G. A. GUTMAN. 1985. Cross-hybridizing snake satellite, *Drosophilia*, and mouse DNA sequences may have arisen independently. Mol. Biol. Evol. **2**:494–504.

LI, W.-H., C.-C. LUO, and C.-I. WU. 1985. Evolution of DNA sequences. Pp. 43–54 *in* R. J. MACINTYRE, ed. Molecular evolutionary genetics. Plenum, New York.

LIVNEH, Z. 1983. Directed mutagenesis method for analysis of mutagen specificity: application to ultraviolet-induced mutagenesis. Proc. Natl. Acad. Sci. USA **80**:237–241.

LU, A.-L., S. CLARK, and P. MODRICH. 1983. Methyl-directed repair of DNA base-pair mismatches *in vitro*. Proc. Natl. Acad. Sci. USA **80**:4639–4643.

MACE, H. A. F., H. R. B. PELHAM, and A. A. TRAVERS. 1983. Association of an S_1 nuclease-sensitive structure with short direct repeats 5′ of *Drosophila* heat shock genes. Nature **304**: 555–557.

MIKLOS, G. L. G. 1985. Localized highly repetitive DNA sequences in vertebrate and invertebrate genomes. Pp. 241–321 *in* R. J. MACINTYRE, ed. Molecular evolutionary genetics. Plenum, New York.

MIKLOS, G. L. G., and A. C. GILL. 1981. The DNA sequences of cloned complex satellite DNA's from Hawaiian *Drosophila* and their bearing on satellite DNA sequence conservation. Chromosoma **82**:409–427.

———. 1982. Nucleotide sequences of highly repeated DNAs: compilation and comments. Genet. Res. (Camb.) **39**:1–30.

MOORE, G. P. 1983. Slipped-strand mispairing and the evolution of introns. Trends Biochem. Sci. **8**:411–414.

NAKASEKO, Y., Y. ADACHI, S.-I. FUNAHASHI, O. NIWA, and M. YANAGIDA. 1986. Chromosome

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

NICKOL, J. M., and G. FELSENFELD. 1983. DNA conformation at the 5′ end of the chicken adult β-globin gene. Cell **35**:467–477.

OHNO, S. 1970. Evolution by gene duplication. Springer, New York.

———. 1972. So much "junk" DNA in our genome. Pp. 366–370 *in* H. H. SMITH, ed. Evolution of genetic systems: Brookhaven Symposium, no. 26. Gordnand Breach, New York.

———. 1981. (AGCTG) (AGCTG) (AGCTG) (GGGTG) as the primordial sequence of intergenic spacers: the role in immunoglobulin class switch. Differentiation **18**:65–74.

ORGEL, L. E., and F. H. C. CRICK. 1980. Selfish DNA: the ultimate parasite. Nature **284**:604–607.

OWEN, J. E., D. W. SCHULTZ, A. TAYLOR, and G. R. SMITH. 1983. Nucleotide sequence of the lysozyme gene of bacteriophage T4: analysis of mutations involving repeated sequences. J. Mol. Biol. **165**:229–248.

PRIBNOW, D., D. C. SIGURDSON, L. GOLD, B. S. SINGER, C. NAPOLI, J. BROSIUS, T. J. DULL, and H. F. NOLLER. 1981. rII cistrons of bacteriophage T4: DNA sequence around the intercistronic divide and positions of genetic landmarks. J. Mol. Biol. **149**:337–376.

RADDING, C. M. 1978. Genetic recombination: strand transfer and mismatch repair. Annu. Rev. Biochem. **47**:847–880.

RAZIN, A., and A. D. RIGGS. 1980. DNA methylation and gene function. Science **210**:604–610.

RIPLEY, L. S. 1982. Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. Proc. Natl. Acad. Sci. USA **79**:4128–4132.

RODAKIS, G. C., and F. C. KAFATOS. 1982. Origin of evolutionary novelty in proteins: how a high-cysteine chorion protein has evolved. Proc. Natl. Acad. Sci. USA **79**:3551–3555.

RODAKIS, G. C., R. LECANIDOU, and T. H. EICKBUSH. 1984. Diversity in a chorion multigene family created by tandem duplications and a putative gene-conversion event. J. Mol. Evol. **20**:265–273.

ROGERS, J. 1983. CACA sequences—the ends and the means? Nature **305**:101–102.

ROSENBERG, H., M. F. SINGER, and M. ROSENBERG. 1978. Highly reiterated sequences of SIMIANSIMIANSIMIANSIMIANSIMIAN. Science **200**:394–402.

SAKOYAMA, Y., Y. YAOITA, and T. HONJO. 1982. Immunoglobulin switch region-like sequences in *Drosophila melanogaster*. Nucleic Acids Res. **10**:4203–4214.

SCHMID, C. W., and C.-K. J. SHEN. 1985. *In* R. J. MACINTYRE, ed. Molecular evolutionary genetics. Plenum, New York.

SINGER, M. F. 1982. Highly repeated sequences in mammalian genomes. Int. Rev. Cytol. **76**: 67–112.

SKOWRONSKI, J., A. PLUCIENNICZAK, A. BEDNAREK, and J. JAWORSKI. 1984. Bovine 1.709 satellite: recombination hotspots and dispersed repeated sequences. J. Mol. Biol. **177**:399–416.

SLIGHTOM, J. L., A. E. BLECHL, and O. SMITHIES. 1980. Human fetal G_γ- and A_γ-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell **21**:627–638.

SMITH, G. P. 1973. Unequal crossover and the evolution of multigene families. Cold Spring Harbor Symp. Quant. Biol. **38**:507–513.

———. 1976. Evolution of repeated DNA sequences by unequal crossing-over. Science **191**: 528–535.

SORGE, J., and S. H. HUGHS. 1982. Polypurine tract adjacent to the U3 region of the Rous sarcoma virus genome provides a cis-acting function. J. Virol. **43**:482–488.

SOUTHERN, E. M. 1975. Long range periodicities in mouse satellite DNA. J. Mol. Biol. **94**:51–69.

SPRITZ, R. A., J. K. DERIEL, B. G. FORGET, and S. M. WEISSMAN. 1980. Complete nucleotide sequence of the human δ-globin gene. Cell **21**:639–646.

STRAUS, N. A., and H. C. BIRNBOIM. 1976. Polypyrimidine sequences found in eukaryotic DNA have been conserved during evolution. Biochim. Biophys. Acta **454**:419–428.

STREISINGER, G., Y. OKADA, J. EMRICH, J. NEWTON, A. TSUGITA, E. TERZHAGHI, and M. INOUYE. 1966. Frameshift mutations and the genetic code. Cold Spring Harbor Symp. Quant. Biol. **31**:77–84.

STREISINGER, G., and J. OWEN. 1985. Mechanisms of spontaneous and induced frameshift mutation in bacteriophage T4. Genetics **109**:633–659.

STRICKBERGER, M. W. 1985. *In* Genetics. 3d ed. Macmillan, New York.

TAUTZ, D., and M. RENZ. 1984*a*. Simple DNA sequences of *Drosophila virilis* isolated by screening with RNA. J. Mol. Biol. **172**:229–235.

———. 1984*b*. Simple sequences are ubiquitous components of eukaryotic genomes. Nucleic Acids Res. **12**:4127–4138.

TAUTZ, D., M. TRICK, and G. A. DOVER. 1986. Cryptic simplicity in DNA is a major source of genetic variation. Nature **322**:652–656.

TOPAL, M. D., and J. R. FRESCO. 1976. Complementary base pairing and the origin of substitution mutations. Nature **263**:285–289.

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

WARNER, R. C., R. A. FISHEL, and F. C. WHEELER. 1979. Branch migration in recombination. Cold Spring Harbor Symp. Quant. Biol. **43**:957–968.

WEINTRAUB, H. 1983. A dominant role for DNA secondary structure in forming hypersensitive structures in chromatin. Cell **32**:1191–1203.

WELLS, R. D., H. BUCHI, H. KOSSEL, E. OHTSUKA, and H. G. KHORANA. 1967*a*. Studies on polynucleotides. LXX. Synthetic deoxyribopolynucleotides as templates for the DNA polymerase of *Escherichia coli:* DNA-like polymers containing repeated tetranucleotide sequences. J. Mol. Biol. **27**:265–272.

WELLS, R. D., T. M. JACOBS, S. A. NARANG, and H. G. KHORANA. 1967*b*. Studies on polynucleotides. LXIX. Synthetic deoxyribopolynucleotides as templates for the DNA polymerase of *Escherichia coli:* DNA-like polymers containing repeating trinucleotide sequences. J. Mol. Biol. **27**:237–263.

WILDEMAN, I. R., and R. N. NAZAR. 1986. A "CAT" family of repetitive DNA sequences in *Saccharomyces cervisiae.* J. Biol. Chem. **261**:13401–13403.

WILLARD, C., E. WONG, J. F. HESS, C.-K. J. SHEN, B. CHAPMAN, A. C. WILSON, and C. W. SCHMID. 1985. Comparison of human and chimpanzee ζ 1 globin genes. J. Mol. Evol. **22**: 309–315.

WALTER M. FITCH, reviewing editor

Received April 23, 1986; revision received November 5, 1986.