Tandem-repetitive Noncoding DNA: Forms and Forces[1]

Wolfgang Stephan

Laboratory of Molecular Genetics, National Institutes of Environmental Health Sciences

A model of sequence-dependent, unequal crossing-over and gene amplification (slippage replication) has been simulated in order to account for various structural features of tandemly repeated DNA sequences. It is shown that DNA whose sequence is not maintained by natural selection will exhibit repetitive patterns over a *wide* range of recombination rates as a result of the interaction of unequal crossing-over and slippage replication, processes that depend on sequence similarity. At high crossing-over frequencies, the nucleotide patterns generated in the simulations are simple and highly regular, with short, nearly identical sequences repeated in tandem. Decreasing recombination rates increase the tendency to longer and more-complex repeat units. Periodicities have been observed down to very low recombination rates (one or more orders of magnitude lower than mutation rate). At such low rates, most of the sequences contain repeats which have an extensive substructure and a high degree of heterogeneity among each other; often higher-order structures are superimposed on a tandem array. These results are compared with various structural properties of tandemly repeated DNAs known from eukaryotes, the spectrum ranging from simple-sequence DNAs, particularly the *hypervariable* minisatellites, to the classical satellite DNAs, located in chromosomal regions of *low* recombination, e.g., heterochromatin.

Introduction

Smith (1976) proposed that DNA whose sequence is not maintained by natural selection will exhibit tandem-repetitive patterns as a result of unequal crossing-over, if strand exchange in the recombinational process is assumed to depend on sequence similarity. He demonstrated this phenomenon by simulating a model of unequal crossing-over for two different rates of crossing-over. Both rates were chosen from a relatively narrow parameter range, such that both are high relative to the mutation rate.

In the past decade, however, tandemly repeated, noncoding DNA sequences have been localized in regions of the eukaryotic genome that are associated with unexpectedly different levels of recombination. Tandem arrays of simple DNA sequences—like the human minisatellites (Jeffreys et al. 1985, 1988)—are thought to be associated with chromosomal regions of high recombination. In contrast, satellite DNAs are located in regions of suppressed recombination (heterochromatin) (Miklos 1985). These data challenge the validity of Smith's hypothesis which has been demonstrated only for relatively high values of the recombination rate. In particular, the evidence that highly

[1] Key words: Sequence-dependent unequal crossing-over, slippage replication, satellite DNA, simple-sequence DNA, minisatellites.

Address for correspondence and reprints: Dr. Wolfgang Stephan, Laboratory of Molecular Genetics, National Institutes of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, North Carolina 27709.

Mol. Biol. Evol. 6(2):198–212. 1989.

0737-4038/89/0602-0007$02.00

repeated DNA sequences (HRDNA), such as satellite DNAs, can accumulate in chromosomal regions of virtually zero crossing-over makes a reconsideration of Smith's analysis necessary.

There is a distinct need for quantitative analysis of the accumulation of noncoding tandem-repetitive sequences. The dependence of various properties of evolving tandem arrays on mechanistic parameters should be evaluated. The quantitative studies conducted so far focus on the lengths of tandem arrays rather than on their structures (Ohta and Kimura 1981; Ohta 1983; Charlesworth et al. 1986; Stephan 1986, 1987; Walsh 1987*a*). In the present paper, I concentrate on the structural patterns of tandem-repetitive sequences as well as on the evolutionary forces creating and maintaining them. The main purpose of my paper is to show how the structure of an array changes with varying recombination rates. By *structure* I mean the length and complexity (substructure) of a repeat unit, the degree of divergence between different repeats of an array, and higher-order periodicities superimposed on an underlying tandem array.

To describe the structural patterns of tandem-repetitive DNA families, located along the chromosome arms in regions that are associated with a wide range of recombination values, it is necessary to extend Smith's analysis along the lines of our previous population genetics model of the evolution of tandemly repeated DNA sequences (Charlesworth et al. 1986; Stephan 1987). That is, the DNA under consideration is assumed to be functionless, and therefore is selectively neutral, when the sequence is short. When it grows too long, it becomes deleterious. In contrast, Smith assumed that, while the sequence itself is not maintained by selection, its length is kept under stabilizing selection, such that the fitness values assigned were 1 when sequence length was in the interval $[A, \Omega]$, where $A > 1$, and 0 elsewhere. Although the above modification of Smith's selection scheme is not essential for generating periodicities per se, it is important because it requires an explicit mechanism for gene amplification to be introduced in order to obtain a nontrivial equilibrium distribution of sequence lengths [see the discussion of this problem in Walsh (1987*a*) and Stephan (1987)]. While studying a model of unequal crossing-over comparable with Smith's, I have examined two sorts of amplification mechanisms: (1) a random amplification model in which a random segment of the sequence is duplicated in tandem and (2) a sequence-dependent amplification mechanism, mimicking replication slippage. A comparison of both models shows that sequence-dependent, unequal crossing-over is capable of creating nucleotide periodicities down to very low recombination rates, when it is accompanied by an amplification mechanism that is also sequence dependent. On the other hand, a random amplification process (in combination with unequal crossing-over) may suffice to describe the structural properties of tandem arrays observed at high recombination rates.

The present paper is organized as follows: In the next section, the model is formulated in detail. That section is followed by a brief description of the simulation process. It includes a list of the basic parameters of the model and the definitions of higher-order structures necessary to analyze the simulation data. Results and discussion are given in the last two sections.

The Model

In a given generation, the following processes are allowed to modify a string of nucleotides in sequence and length: base substitutions, amplification, unequal sister-chromatid exchange (SCE), and natural selection. I consider only sister-strand exchanges. Simulating only SCE uses many fewer computer resources, since, instead of keeping track of an entire population, only a single chromosome lineage is followed through time. The processes that are assumed to depend on sequence similarity are described first.

Unequal Sister-Strand Exchange

SCE is modeled as a two-step process, including (1) misalignment of the chromatids and (2) crossing-over at a certain position in the region of strand overlap. Two identical strands are aligned according to a probability distribution $P_i(k)$, where i denotes the length of the nucleotide string and k denotes the number of nucleotides by which the strings are aligned out of register. Since the molecular basis for misalignment of sister chromatids is largely unknown, I express $P_i(k)$ simply by the identity coefficients, $f_i(k)$, used by Ohta (1980), as follows:

$$P_i(k) = f_i(k) / \sum_k f_i(k) , \tag{1}$$

where $f_i(k)$ is the probability that two nucleotides that are k steps apart are identical and where $1 \leq k \leq d$, with $d \leq$ (sequence length $-$ 1). d denotes the maximum number of nucleotides two sequences are allowed to be out of register. Rough estimates of d are known from satellite DNA and minisatellites in humans. The largest (unequal) exchanges reported for five different minisatellite families amount to ~20% of the total array length (Jeffreys et al. 1988). Large changes in the amount of heterochromatin have been observed in humans between parental and daughter chromosomes, suggesting that unequal crossing-over can involve a large number of repeats of satellite DNA (Craig-Holmes et al. 1975).

While misalignment is assumed to be a function of the overall sequence similarity (note that the identity coefficients are quantities averaged over the entire sequence overlap), the position and frequency of crossing-over are determined by local properties. These concern the recognition and synapsis of complementary single DNA strands. In Smith's model of unequal exchange, and likewise here, the central quantity characterizing the synapsis reaction of two complementary strands is the match length, m, i.e., the minimal number of bases of perfect match that are required to obtain stable pairing. The phenomenological parameter m is directly measurable (Ayares et al. 1986; Walsh 1987*b*).

Sequence similarity is an essential component of all proposed mechanisms of recombination (Holliday 1964; Meselson and Radding 1975; Szostak et al. 1983). For formulation, I use here the Holliday-Meselson-Radding scheme of general recombination, which allows only for single-strand breakage. In general recombination, strand breakage can occur at any position in the region of chromatid overlap, and the enzymes involved can use either pair of sequences as substrates. The breakage allows movement of the ends created by the nicks to the other duplex strand. A connection can be created when one of the ends has at least m nucleotides in common with the target strand at the position of strand breakage. To formulate this process, consider two identical single DNA strands which are aligned according to distribution $P_i(k)$. In the first step, all sequences in the overlap region that match perfectly and are exactly m nucleotides long are identified. When the strings have a stretch of $\hat{m}$ ($\geq m$) nucleotides in common, $\hat{m} - m + 1$ sequences of length m are marked within this stretch, since each of them provides a possibility for synapsis of the strands. If n_m is the resulting number of possible attachment sites of length m and if $\tilde{\gamma}_m$ is the rate of crossing-over for two sequences with match length m, the frequency of recombination in the entire overlap region is given by a Poisson distribution with mean $n_m\tilde{\gamma}_m$ and its position is given by the distribution of all possible attachment sites marked. Once such a site of length m is determined, strand breakage occurs at random at its 3′ or 5′ end.

It follows from these considerations that when two strands share a stretch of $\hat{m}$ ($\geq m$) bases of perfect match, the frequency of crossing-over within this region is given by

$$(\hat{m} - m + 1)\tilde{\gamma}_m \, . \tag{2}$$

Such a linear relationship has been found for intermolecular recombination in mammals over a wide range of $\hat{m}$ (Ayares et al. 1986). $\tilde{\gamma}_m$, the frequency of crossing-over for sequences having a stretch of the minimal match length m in common and hence the unit of the recombination rate, is considered an adjustable parameter. It can vary along the chromosome arms for several reasons, for instance because it depends on the chromatin structure or on the protein(s) mediating strand exchange. It is at present unclear whether and to what extent this quantity depends on the match length m. In a first approximation, I consider therefore $\tilde{\gamma}_m$ an adjustable parameter that is independent of m; that is,

$$\tilde{\gamma}_m = \tilde{\gamma} \, . \tag{3}$$

Replication Slippage

Under the processes leading to an average increase in array length (called here *amplification*) are some which depend on sequence similarity, at least in certain stages. I give here an explicit model for replication slippage that is likely to be the most important amplification mechanism for simple-sequence DNA (Levinson and Gutman 1987). Whether satellite DNAs are amplified by this mechanism is unclear. The existence of extrachromosomal circular satellite DNAs (see review by Walsh 1987*a*) suggests an extrachromosomal mode of amplification (e.g., rolling circle; Hourcade et al. 1973). Sequence similarity required for the reinsertion of the overreplicated DNA into the genome appears to be an important component of this process, as it is for replication slippage. Because of this similarity requirement, the role of a rolling circle–type mechanism in the formation of tandem-repetitive sequences can be expected to be the same as that of slippage.

Slippage during DNA replication occurs when parental and the newly synthesized daughter strand occasionally unwind and subsequently mispair, a process resulting in duplications or deletions of sequences. In the present paper, I assume that slippage is biased toward amplification such that certain sections of the sequence will be overreplicated. For formulation, consider a string of i nucleotides and choose an integer i_P at random, such that $1 \leq i_P \leq i$. Here, 1 corresponds to the 3′ end of the strand and i_P denotes the coordinate of the point to which the DNA polymerase has proceeded on the template strand moving 3′ to 5′, when the newly synthesized strand melts off from the template. Such an event is assumed to happen at rate μ_s per generation. Rebinding is allowed only at those sites of the already utilized template where the free end of the new strand finds a stretch of m_s exactly complementary bases. This is the crucial step in the replication process that makes it dependent on sequence similarity, in analogy to the synapsis reaction in recombination.

Realization of replication slippage on the computer is similar to the case of SCE. All unpaired sites between the 3′ end and point i_P where the end of the new strand can attach are marked, if they are located within a certain distance, d_s, to i_P. This constraint is introduced because slippage is thought to involve only a limited number of nucleotides (Levinson and Gutman 1987). The site where attachment eventually occurs is determined at random from the ensemble of all possible binding sites. If i_P does not coincide with the 5′ end of this site, the sequence lying between the 5′ end and i_P is replicated twice, so that the newly synthesized strand contains a tandemly duplicated region. The nucleotide string complementary to the new strand is taken to continue the simulations.

Natural Selection

The effects of natural selection of individuals carrying tandemly arrayed, noncoding sequences in their genomes are taken into account in the following way: I assume that the sequence is not maintained by selection but that the length of the sequence is under selective constraints. I apply a truncation selection scheme, as described elsewhere (Stephan 1987), assuming that simple-sequence and satellite DNAs are neutral when present in small amounts and are deleterious at too high amounts. If Ω denotes the upper limit of the total length of the sequence tolerable for the organisms, the dependence of the fitness function, w_i, on sequence length, i, is chosen as

$$w_i = \begin{cases} 1, & i \leqslant \Omega \\ 0, & i > \Omega \,. \end{cases} \tag{4}$$

Simulation Process and Definitions

The Simulations

In the computer simulations, a single chromosome lineage is followed through time. The general process occurs in four steps. At the beginning of each generation a mutated strand is produced by base substitutions, such that each substitution is equally likely. In the second step, the strand may undergo amplification of certain segments in one of two ways. The first way is random amplification in which a randomly chosen segment of the sequence is tandemly duplicated. The second way is by replication slippage. In the third step, the resulting strand may undergo SCE. In the fourth step, one of the products is chosen for the next generation. The chromosome continuing the line in the next generation is chosen at random, unless one of the sequences exceeds the selection boundary Ω as a result of SCE or amplification. In this case the smaller one is taken to start the next cycle.

This process requires the following parameters: The rate (per generation) at which a nucleotide is substituted in step 1 is $\tilde{u}$. The minimum number of bases that must match before sister-strand exchange (replication slippage) can occur is m (m_s). The rate at which two sequences with m (m_s) consecutive matches will undergo sister-strand exchange (replication slippage) is $\tilde{\gamma}(\mu_s)$. The maximum distance over which the search is made for m (m_s) matching bases is d (d_s). In random amplification, μ_s is replaced by μ, the rate of duplication per generation. The maximum length of the sequence which may undergo all those processes is Ω.

Since the processes of mutation, SCE, and amplification are assumed to occur independently of one another, the results can be expressed as functions of the relative rates of these processes, i.e., $\tilde{\gamma}/\tilde{u}$ and $\tilde{\gamma}/\mu$ (or $\tilde{\gamma}/\mu_s$). In the simulations, the value of $\tilde{\gamma}/\tilde{u}$ has been varied by varying $\tilde{\gamma}$, while keeping $\tilde{u}$ constant. Thus, a phrase such as "high (low) recombination rate," frequently used in the next two sections, is synonymous with "high (low) values of $\tilde{\gamma}/\tilde{u}$."

Initiation and termination of the simulations were done in the following way: Each simulation run was started with a string of 10 identical nucleotides. Simulation cycles were performed until statistical equilibrium was reached. According to the usual criterion for statistical equilibrium (Ohta 1980), the simulations were terminated after the $(2/v)$-th generation, where v is the rate constant of the rate-limiting step of the process. Application of this criterion is straightforward, when mutation is much slower than SCE. In this case, $v = \tilde{u}$. However, at low recombination rates, $\tilde{\gamma}$, one has to take into account that SCE depends on sequence similarity and that the effective rate of recombination can be much lower than $\tilde{\gamma}$, when sequence divergence is sufficiently high. For instance, under the assumption that sequence differences should be <50% for repetitive structures to be recognizable, the simulations were run 50 times longer in the case $\tilde{\gamma} = \tilde{u}$ than was required by the above criterion. This factor is estimated using equations (5a) and (5b) of the next section. For each parameter set I made 10 replicates.

Definition of Higher-Order Structures

I call the repeating element of a tandem-repetitive sequence R. $(R)_n$ denotes a tandem array consisting of n such tandem elements R. It may occur that the repeat unit R has a recognizable internal structure. If this internal structure is imperfect, I call it a *substructure.* Various sorts of substructures can arise (see below). For instance, the repeat units of a tandem array may contain runs of consecutive identical nucleotides (called *homopolymeric tracts*) and dinucleotides that are interrupted by other interposed sequences. However, one may observe that subsets of the R's are more closely related to each other than to the other subset(s) of R in a very regular way, such that (*a*) subset consensus sequences (see definition in table 2A) of equal lengths exist and (*b*) the arrangement of the subsets in all R's is identical. In this case, the subset is considered the repeat unit and R is called a *higher-order periodicity.* The simplest higher-order periodicity is a dimer. That means R can be represented in the form R_1R_2, where R_1 and R_2 are subsets of R with equal lengths, such that the similarity within the R_1 and R_2 elements (across the tandem array) is much greater than that between the R_1 and R_2 consensus sequences. In this example, the tandem array $(R)_n$ can be written as $(R_1R_2)_n$. Alternatively, one may observe the formation of distinct domains within a tandem array, in the simplest form, as $(R_1)_k(R_2)_m$, where k and m are the copy numbers of the R_1- and R_2-repeating elements, respectively. In this case, the repeats R_1 and R_2 need not have equal lengths. I refer to such higher-order structures as *multiple arrays.*

Results

Interaction of Sequence-dependent SCE and Random Amplification

Table 1 shows two typical repetitive sequences generated in the simulations. They were obtained for high recombination rates (relative to the mutation rate) comparable with those chosen by Smith (1976, table 1). At those high values, 8 of 10 simulations led to repetitive patterns; one sequence was homopolymeric and another one heter-

Table 1
Examples of Repetitive Sequences Generated by Sequence-dependent Unequal Crossing-over and Random Amplification at a High Recombination Rate

...... GATAGAGAGAGAGAGCGAGAGAGAGAGACGAGAGAGGAGAGAGAGAGAGAGGAGAGAGA ......
...... GATGTGATGTGATGTGAGATGTGATGTGATGTGATGTGATGAGATGTGATGT ......

NOTE.—These nucleotide tracts consist of tandem-repetitive sequences (shown) and irregular flanking regions (indicated by dots). The repeat units are GA and GATGT, respectively. Underlining indicates imperfections in the repetitive regions. The parameter values are $\tilde{\gamma}/\tilde{u} = 125$; $\mu/\tilde{\gamma} = 1$; $m = 5$ bp; and $\Omega = 1{,}000$ bp. Unequal crossing-over is not restricted; that is, d = (sequence length − 1).

ogeneous. For the repetitive nucleotide tracts, two features are common. First, they are composed of both a sequence showing a repetitive structure and irregular flanking regions (indicated by dots in table 1). This pattern can be interpreted as follows: Sequence-dependent unequal crossing-over is capable of generating periodicities in genomic regions, in which exchanges occur frequently enough relative to mutation or amplification events. This is the case in the repetitive region, since crossover positions are located preferentially in the middle of a sequence because of similarity requirements. Second, the repetitive region is sometimes interrupted by interposed single nucleotides or short sequences. This can be attributed to the ongoing mutation process but is here mainly caused by random amplification events, since $\mu \gg \tilde{u}$. Random amplification works in this model as an additional mutational force, generating "noise" which is then counteracted by SCE, restoring the pattern. For instance, for the sequences in table 1 only a few rounds of SCE would be needed to remove the irregularities, if no new ones were produced. These examples may suffice to demonstrate the ability of sequence-dependent, unequal crossing-over to initiate and maintain repetitive structures in nucleotide strings at high recombination rates. Smith (1976) obtained similar results for such high recombination values.

The effects of unequal crossing-over on the formation of nucleotide periodicities have been obtained by varying the recombination rate $\tilde{\gamma}$, while keeping mutation rate $\tilde{u}$ constant. The results are displayed in table 2 for three different values of $\tilde{\gamma}/\tilde{u}$. At high recombination rates, repeats are generally very short, with di- and trinucleotides being predominant. In contrast, decreasing crossing-over frequencies led to longer repeat units, and, in parallel, more simulations produced nonrepetitive structures. For $\tilde{\gamma}/\tilde{u} = 125$, 8 of 10 runs led to periodic patterns, as mentioned above, while for $\tilde{\gamma}/\tilde{u} = 25$ and 5, periodicities were produced in only 6 and 1 runs, respectively, of 10 (see table 2). For low recombination rates ($\tilde{\gamma}/\tilde{u} = 1$) periodicity disappeared completely. Twenty simulations were run for this value of $\tilde{\gamma}/\tilde{u}$ and $m = 5$, but none of them led to recognizable periodicities, and the results for even smaller match lengths ($m = 2$ or 3) were also negative.

Match length has an inverse effect on repeat length, as indicated in table 3 for three different values of m (while recombination rate is fixed). For $m = 3$, all 10 simulations led to periodicities. Their repeat units are very short. For increasing m, there is a clear tendency to longer repeating units, and, as a consequence, fewer nucleotide strings generated in the simulations exhibit a repetitive pattern. These results are in accordance with those above on varying recombination levels, in that increasing match lengths, like decreasing recombination rates, lower the frequency of exchange relative to mutation and random amplification and hence reduce the efficiency of unequal crossing-over in generating periodic patterns.

Table 2
Repeat Length as a Function of $\tilde{\gamma}/\tilde{u}$ for Sequence-dependent Unequal Crossing-over and Random Amplification

$\tilde{\gamma}/\tilde{u}$		
125	25	5
2	18	52
2	25	. . .
2	26	. . .
3	28	. . .
3	43	. . .
5	62	. . .
7	. . .	. . .
24	. . .	. . .

NOTE.—For each set of parameters, 10 simulations were run. The table displays only the results of those runs that produced periodic structures. For $\tilde{\gamma}/\tilde{u}$ = 125, 8 of 10 simulations led to periodicities, whereas, for the lower recombination rates $\tilde{\gamma}/\tilde{u}$ = 25 and 5, repetitive patterns were generated in only 6 and 1 runs, respectively. In parallel, the lengths of the repeat units increase with decreasing recombination frequencies. The values of the parameters other than $\tilde{\gamma}/\tilde{u}$ are fixed as in table 1.

Interaction of Sequence-dependent SCE and Slippage Replication

I carried out simulations for various values of $\tilde{\gamma}/\tilde{u}$ and m. Some of the results are displayed in the tables 4–6. The overall effects of these parameters on repeat length are the same as for random amplification. That is, decreasing recombination rates and/or increasing match lengths lead to longer repeat units and reduce the propensity of unequal crossing-over and amplification to form repetitive patterns. However, there is one important difference. One aspect of this is the fact that, at high recombination rates ($\tilde{\gamma}/\tilde{u}$ = 125), repetitive structures are totally regular, having no interposed sequences, unlike in the case of random amplification (table 1) (data not shown).

Another aspect of this phenomenon is that sequence-dependent unequal exchange, in concert with slippage replication, is capable of generating periodicities down to very low recombination rates (table 6). The reason for that is obvious. An amplification mechanism that depends on sequence similarity does not destroy preexisting patterns, as is the case when random sections of the genome are duplicated. Instead, amplification reinforces unequal crossing-over in maintaining and even creating periodicity, because both mechanisms are coupled via similarity requirements. This synchronization phenomenon can be demonstrated directly, if it is assumed that the match lengths of SCE and slippage are different, e.g., $m = 5$ and $m_s = 3$. In this case, slippage replication and unequal crossing-over cannot completely synchronize. That means the tandem-repetitive structures generated in the simulations are determined partly by the process with the larger match length (in this case, unequal crossing-over) and partly by the other process. A typical pattern I observed at high recombination rates was as follows: A tandem array consisting of relatively long repeat units was followed by a tandem array with fewer copies of a smaller repeating element (often a subrepeat of the other one), followed again by the array with the longer repeats, and so on. This phenomenon

Table 3
Repeat Length as a Function of *m* for Sequence-dependent Unequal Crossing-over and Random Amplification

m		
3	5	10
2	2	5
2	2	5
2	2	12
2	3	14
3	3	20
3	5	22
3	7	. . .
5	24	. . .
5	. . .	. . .
5	. . .	. . .

NOTE.—An increasing match length *m* leads to longer repeating units. Consequently, repetitive structures arise less often. Apart from *m*, the parameters were chosen as in table 1.

Table 4
Structures of Tandem-repetitive Sequences Generated by Sequence-dependent Unequal Crossing-over and Replication Slippage for $\tilde{\gamma}/\tilde{u}$ = 25

REPEAT LENGTH (bp)		HETEROGENEITY (%)			
A	B	A	B	AB	STRUCTURE
3		0.00			Monomers: $(A)_{13}$
6		3.85			Monomers: $(A)_{13}$
8		8.33			Monomers: $(A)_6$
12		0.33			Monomers: $(A)_{25}$
5	5	4.29	3.14	40.00	Dimers: $(AB)_7$
11	11	1.51	0.00	9.09	Trimers: $(AB_2)_6$
8	8	0.00	0.00	25.00	Double array: $(A)_7(B)_6$
8	8	0.00	0.00	25.00	Double array: $(A)_4(B)_{10}$
15	15	0.00	4.44	26.67	Double array: $(A)_8(B)_{10}$
30	30	0.00	1.11	10.00	Double array: $(A)_7(B)_{18}$

NOTE.—In those cases in which the tandem repeats were heterogeneous, the consensus sequence was taken as the repeating unit. This is the sequence representing the most abundant nucleotide at each position in the repeats comprising an array. For the given parameter set, all 10 simulations led to repetitive structures. As indicated in the last column, the first four examples of the table are simple repetitions of a short sequence, the next two examples show higher-order periodicities (dimers $(AB)_n$ and trimers $(AB_2)_n$), and the last four sequences generated are composed of two distinct tandem arrays of the form $(A)_k(B)_n$, where A and B are the respective consensus sequences and *k* and *n* are their copy numbers. The repeat lengths of A and B are given in the first and second columns, respectively. The degrees of heterogeneity among the A's and B's within an array are shown in the next two columns, while the fifth column gives the heterogeneity between the A- and B-consensus sequences. The parameter values are: $\mu_s/\tilde{\gamma} = 10$; $m = m_s = 5$ bp; $\Omega = 1{,}000$ bp; and $d_s = 50$ bp. Unequal crossing-over is restricted such that chromatid overlap is ⩾40 bp.

Table 5
Structures of Tandem-repetitive Sequences Generated by Sequence-dependent Unequal Crossing-over and Replication Slippage for $\tilde{\gamma}/\tilde{u}$ = 5

Repeat Length (bp)		Heterogeneity (%)			
A	B	A	B	AB	Structure
9		5.80			Monomers: $(A)_{55}$
14		0.56			Monomers: $(A)_{64}$
14		13.60			Monomers: $(A)_{24}$
18		1.85			Monomers: $(A)_{9}$
18		6.12			Monomers: $(A)_{10}$
20		0.00			Monomers: $(A)_{13}$
22		0.89			Monomers: $(A)_{41}$
12	12	0.64	6.09	58.33	Dimers: $(AB)_{26}$
14	14	3.57	7.14	28.57	Dimers: $(AB)_{5}$
9	9	11.11	12.34	55.56	Pentamers: $(AB_4)_{11}$

NOTE.—All 10 simulation runs produced periodic structures. Apart from $\tilde{\gamma}/\tilde{u}$, the parameters were chosen as in table 4.

is analogous to "frequency pulling" of coupled, nonlinear oscillators (Minorsky 1962). Two out-of-step oscillators tend to synchronize, if their natural frequencies are close to one another, but synchronization is incomplete, if the frequencies are too far apart.

Discussion

Periodicities Exist at Low Recombination Rates

Simulating a model of unequal crossing-over and slippage replication, I observed periodicities in 5 of 10 runs, when $\tilde{\gamma}/\tilde{u}$ was as low as 1 (table 6). At this low level of crossing-over, no periodic structures have been found for the model of unequal exchange and random amplification. This has some important bearings on previous

Table 6
Structures of Tandem-repetitive Sequences Generated by Sequence-dependent Unequal Crossing-over and Replication Slippage for $\tilde{\gamma}/\tilde{u}$ = 1

Repeat Length (bp)		Heterogeneity (%)			
A	B	A	B	AB	Structure
20		20.10			Monomers: $(A)_{67}$
33		8.08			Monomers: $(A)_{18}$
85		21.17			Monomers: $(A)_{22}$
9	9	6.06	5.55	66.67	Double array: $(A)_{18}(B)_{53}$

NOTE.—For this set of parameter values, only 5 of 10 simulations led to periodicities. In addition to the examples given in the table, a triple array of the form $(A)_{25}(B)_{74}(C)_{44}$ was generated. The lengths of A, B, and C were 17 bp, and their degrees of heterogeneity were 34.12%, 41.96%, and 31.37%, respectively, i.e., relatively high. The heterogeneity between the A- and B-consensus sequences was 70.59%, 52.94% between A and C, and 82.35% between B and C. To account for the possible occurrence of long repeat units and high degrees of sequence heterogeneity at this low recombination rate, the total sequence length was chosen as Ω = 5,000 bp and slippage was not restricted.

results on the accumulation of HRDNA (Stephan 1987). To compare these results, one needs to calculate an effective recombination rate, $\hat{\gamma}_{\mathrm{eff}}$, per repeat unit, taking the (sometimes considerable) sequence divergence into account (see table 6). If p is the degree of divergence, the probability of finding a perfect match of length m becomes $(1 - p)^m$. Then, using equation (2), one obtains

$$\hat{\gamma}_{\mathrm{eff}} = \tilde{\gamma}(\hat{m} - m + 1)(1 - p)^m , \tag{5a}$$

where $\hat{m}$ equals the repeat length. For long repeats ($\hat{m} \gg m$), equation (5a) can be approximated in our example ($\tilde{\gamma} = \tilde{u}$) as follows:

$$\hat{\gamma}_{\mathrm{eff}} = \hat{u}(1 - p)^m , \tag{5b}$$

where $\hat{u}$ is the mutation rate per repeat unit. Evaluating this relation for $m = 5$ and $p = 1/3$ and $1/2$ (compare table 6), one finds $\hat{\gamma}_{\mathrm{eff}} = 0.124\hat{u}$ and $0.031\hat{u}$, respectively. That means periodic patterns can be observed, even when the recombination rate is an order of magnitude or more less than the mutation rate, a somewhat surprising result. However, this result is consistent with previous calculations on the accumulation of HRDNA [relation (15) of Stephan (1987)]. The importance of this agreement lies in the fact that the present results have been obtained from specific molecular models of unequal exchange and gene amplification, both formulated in terms of the DNA strand itself, while in my previous model the repeating units were considered abstract copies.

In our previous work on the accumulation of tandemly repeated sequences (Charlesworth et al. 1986; Stephan 1986, 1987), unequal crossing-over could occur either at the premeiotic mitoses or at meiosis itself. At low crossing-over frequencies, the population genetic consequences of meiotic and mitotic exchanges turned out to be essentially identical (Stephan 1986). Our population genetics model was proposed to account for the accumulation of satellite DNA in chromosomal regions of restricted recombination (heterochromatin) and to explain the extreme length heterogeneity of satellite tracts that is frequently found between closely related species but that is only rarely found within a species. There is evidence from various species that mitotic as well as meiotic exchanges are suppressed in heterochromatin (Holmquist and Comings 1975; Szauter 1984), but their relative rates and hence their importance for the evolutionary dynamics of satellite DNAs are still a matter of debate.

Sequence Similarity Requirements for Recombination

Attempts have been made recently to determine for various recombinational processes the degree of sequence similarity required for synapsis of single DNA strands. Most of the known data are from prokaryotes. For eukaryotes, two instances are documented. Ayares et al. (1986) observed for intermolecular recombination in mammalian cells that a stretch of 25 bp of perfect match is sufficient to yield recombinant products. Rubnitz and Subramani (1984) investigated the sequence similarity requirements for intramolecular recombination in monkey cells and found $m = 14$ bp. For simple DNA sequences, m might be even lower.

This is in accordance with the suggestion by Thomas (1966) that the target size for recombination should be smaller in less complex genomic structures, and it is also indicated by the simulations. I found that the lengths of homopolymeric nucleotide tracts, occurring as subunits within repeats, never exceed the match length (data not shown). On the basis of this observation, various examples of simple-sequence and satellite DNAs can be examined to obtain an estimate for m. For instance, the three satellite DNAs of *Drosophila virilis,* each consisting of 7-bp repeats, contain AAA tracts, the HS-α satellite from kangaroo rats formed by 6 bp contains a GGG subunit, and, similarly, the basic 9-bp repeat of mouse satellite contains an AAAAA tract (John and Miklos 1979). These examples suggest that m assumes values of ~5 bp.

A possible explanation for the relationship between the match length and the length of a poly(N) subrepeat is as follows: A homopolymeric nucleotide tract that is longer than m can cause irregularities in the formation of tandem-repetitive patterns, since the positions of crossover and slippage are not exactly determined but vary within the poly(N) subrepeat. As a consequence, irregularities may arise. These, however, will be removed by further rounds of sequence-dependent amplification and SCE, until both processes are eventually entrained. Consistent with this interpretation is that I did not observe a similar phenomenon in the case of random amplification.

Implications for Satellite DNA: Sub- and Superstructures of the Repeating Units

Besides the observation that repeats often contain homopolymeric stretches of nucleotides, other types of substructures have been found in the simulations. Repeating units can be composed of variants of one or more basic repeats. For instance, the fifth sequence in table 5, an 18-bp repeat, consists of three variants of a hexanucleotide, and, similarly, the 30-bp repeat in table 4 consists of five variants of NCCNCG. There are various such examples known for satellite DNAs. The most prominent one is the 232-bp repeat of the mouse major satellite, showing extensive internal substructure (Hörz and Altenburger 1981). It can be divided into four segments of alternating 28- and 30-bp units that are related. Furthermore, Hörz and Altenburger have suggested that these major internal repetitions have arisen from a monomer of the form GAAAAANNN.

On the other hand, I also observed higher-order structures, as defined above. In the simulations, such structures arose when recombination rates were sufficiently low. For instance, for $\tilde{\gamma}/\tilde{u} = 25$ I observed four examples in which two distinct tandem arrays were formed within a nucleotide string. Their respective consensus sequences differ by 10%–26.7% (table 4). For $\tilde{\gamma}/\tilde{u} = 1$, in one case even three tandem arrays were formed simultaneously (table 6). In both examples of multiple arrays observed at this low recombination rate, the differences in the consensus sequences amount to >50%. For satellite DNAs, the latter situation is frequently met in nature; that is, in most species, satellite DNA families differ greatly. However, there are also some exceptions. For instance, the three major satellite families in *D. virilis,* all consisting of heptanucleotide repeat units, differ from one another only by a single nucleotide (John and Miklos 1979).

With respect to higher-order periodicities, a well-known example is the primates' alphoid satellite DNA, whose predominant repeat form is dimeric in most species (Shmookler Reis et al. 1985). Divergence between the two monomers within a dimer can be as high as 30%–55%, whereas ~10% divergence is found among dimeric repeats. The same behavior has been observed in the simulations for $\tilde{\gamma}/\tilde{u} = 5$ in two instances (table 5). For the same parameter set, a pentamer of the form AB_4 was also found. In this case, the average divergence between A and B amounts to 55.5%, while A's and B's differ within themselves by only 11% and 12%, respectively.

Implications for Simple-Sequence DNA

At high recombination rates, in particular for $\tilde{\gamma}/\tilde{u} = 125$, unequal crossing-over, in concert with both random and sequence-dependent amplification mechanisms, produced tandem-repetitive structures that show features typically found in studies of simple-sequence DNAs, i.e., short repeat units and lack of higher-order structures. The repeat unit lengths for random amplification are displayed in table 2. For sequence-dependent amplification, the average repeat length was 6.1 bp for $\tilde{\gamma}/\tilde{u} = 125$, and no higher-order periodicities were observed for this parameter value, while, at a five-fold-lower recombination level, repetitive sequences composed of di- and trimeric repeat forms have been found (table 4).

The fact that repetitive sequences of relatively simple structures were produced at high crossing-over rates is consistent with the notion that simple-sequence DNAs are considered hot spots of recombination. Tandem arrays in which the role of unequal crossing-over as the predominant force in the evolutionary dynamics is demonstrated most clearly are the minisatellite regions in vertebrate genomes, e.g., in humans (Jeffreys et al. 1985, 1988). Examining five different human minisatellites with respect to the rates at which they change in array length (number of repeat units) from generation to generation, Jeffreys et al. (1988, table 1) provide evidence that minisatellite regions composed of the shortest repeats "mutate" (probably by unequal crossing-over) most rapidly. This is in agreement with the prediction, inferred from my simulations, that tandem-repetitive sequences that undergo unequal crossing-over more frequently have shorter repeat units. However, more data are needed to confirm this result. Furthermore, it would be interesting to explore minisatellite structural properties other than the lengths of repeat units, e.g., the complexity of repeats and the presence or absence of higher-order periodicities. A more fundamental question to be solved concerns the recombinagenic nature of minisatellites—i.e., whether these repeats contain a χ-like sequence which serves as a recombination signal (Jeffreys et al. 1985) or whether increased rates of recombination in minisatellite regions are due to sequence similarity and the tandem repetition of simple-sequence DNA, as suggested in the present paper.

There is yet another class of simple tandem-repetitive sequences that are ubiquitous components of eukaryotic genomes (reviewed by Levinson and Gutman 1987). These sequences consist of very short repeats, mostly dinucleotides, and are often <100 bp long. A well-known example is the poly(GT) sequence, present in virtually all organisms of the animal kingdom. It has been localized in flanking regions of genes and in nontranscribed spacer regions but not in coding sequences (Morris et al. 1986). However, it is not clear at present to what extent unequal crossing-over is acting on simple repetitive sequences whose total array length is < ~100 bp, because there may be topological constraints on unequal exchange due to the fact that it is an interhelical process. For this class of tandem arrays, replication slippage, an intrahelical mechanism, could be a major factor in the expansion and contraction dynamics (Levinson and Gutman 1987).

Acknowledgments

I am grateful to C. Langley and B. Charlesworth for many discussions and helpful suggestions on the several versions of the manuscript. Furthermore, I had interesting discussions on this subject with C. Giroux, R. Hudson, N. Kaplan, T. Petes, P. Starlinger, and B. Walsh. I thank the National Institute of Environmental Health Sciences for supporting my visit at this institute in summer 1986, when this project was started. I also thank W. Fitch for his suggestions on improving the manuscript.

LITERATURE CITED

AYARES, D., L. CHEKURI, K.-Y. SONG, and R. KUCHERLAPATI. 1986. Sequence homology requirements for intermolecular recombination in mammalian cells. Proc. Natl. Acad. Sci. USA **83**:5199–5203.

CHARLESWORTH, B., C. H. LANGLEY, and W. STEPHAN. 1986. The evolution of restricted recombination and the accumulation of repeated DNA sequences. Genetics **112**:947–962.

CRAIG-HOLMES, A. P., F. B. MOORE, and M. W. SHAW. 1975. Polymorphism of human C-band heterochromatin. II. Family studies with suggestive evidence of somatic crossing over. Am. J. Hum. Genet. **27**:178–189.

HOLLIDAY, R. 1964. A mechanism for gene conversion in fungi. Genet. Res. **5**:282–304.

HOLMQUIST, G. P., and D. E. COMINGS. 1975. Sister chromatid exchange and chromosome organization based on a bromodeoxyuridine Giemsa-C-banding technique (TC-banding). Chromosoma **52**:245–259.

HÖRZ, W., and W. ALTENBURGER. 1981. Nucleotide sequence of mouse satellite DNA. Nucleic Acids Res. **9**:683–696.

HOURCADE, D., D. DRESSLER, and J. WOLFSON. 1973. The amplification of ribosomal RNA genes involving a rolling circle intermediate. Proc. Natl. Acad. Sci. USA **70**:2926–2930.

JEFFREYS, A. J., N. J. ROYLE, V. WILSON, and Z. WONG. 1988. Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. Nature **332**: 278–281.

JEFFREYS, A. J., V. WILSON, and S. L. THEIN. 1985. Hypervariable 'minisatellite' regions in human DNA. Nature **314**:67–73.

JOHN, B., and G. L. G. MIKLOS. 1979. Functional aspects of satellite DNA and heterochromatin. Int. Rev. Cytol. **58**:1–114.

LEVINSON, G., and G. A. GUTMAN. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4**:203–221.

MESELSON, M. S., and C. M. RADDING. 1975. A general model for genetic recombination. Proc. Natl. Acad. Sci. USA **72**:358–361.

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 and London.

MINORSKY, N. 1962. Nonlinear oscillations. Van Nostrand, Princeton, N.J.

MORRIS, J., S. R. KUSHNER, and R. IVARIE. 1986. The simple repeat poly(dT-dG)·poly(dC-dA) common to eukaryotes is absent from eubacteria and archaebacteria and rare in protozoans. Mol. Biol. Evol. **3**:343–355.

OHTA, T. 1980. Evolution and variation of multigene families. Springer, Berlin, Heidelberg, and New York.

———. 1983. Theorectical study on the accumulation of selfish DNA. Genet. Res. **41**:1–15.

OHTA, T., and M. KIMURA. 1981. Some calculations on the amount of selfish DNA. Proc. Natl. Acad. Sci. USA **78**:1129–1132.

RUBNITZ, J., and S. SUBRAMANI. 1984. The minimum amount of homology required for homologous recombination in monkey cells. Mol. Cell. Biol. **4**:2253–2258.

SHMOOKLER REIS, R. J., A. SRIVASTAVA, D. T. BERANEK, and S. GOLDSTEIN. 1985. Human alphoid family of tandemly repeated DNA: sequence of cloned tetrameric fragments and analysis of familial divergence. J. Mol. Biol. **186**:31–41.

SMITH, G. P. 1976. Evolution of repeated DNA sequences by unequal crossover. Science **191**: 528–535.

STEPHAN, W. 1986. Recombination and the evolution of satellite DNA. Genet. Res. **47**:167–174.

———. 1987. Quantitative variation and chromosomal location of satellite DNAs. Genet. Res. **50**:41–52.

SZAUTER, P. 1984. An analysis of regional constraints on exchange in *Drosophila melanogaster* using recombination-defective meiotic mutants. Genetics **106**:45–71.

SZOSTAK, J. W., T. L. ORR-WEAVER, R. J. ROTHSTEIN, and F. W. STAHL. 1983. The double strand break repair model for recombination. Cell **33**:25–35.

THOMAS, C. A. 1966. Recombination of DNA molecules. Prog. Nucleic Acid Res. Mol. Biol. **5**:315–348.

WALSH, J. B. 1987*a*. Persistence of tandem arrays: implications for satellite and simple sequence DNAs. Genetics **115**:553–567.

———. 1987*b*. Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? Genetics **117**:543–557.

WALTER M. FITCH, reviewing editor

Received December 24, 1987; revision received October 7, 1988