Possible Role of Natural Selection in the Formation of Tandem-Repetitive Noncoding DNA

Wolfgang Stephan* and Soowon Cho†

**Department of Zoology and †Department of Entomology, University of Maryland, College Park, Maryland 20742*

Manuscript received April 24, 1993
Accepted for publication September 7, 1993

ABSTRACT

A simulation model of sequence-dependent amplification, unequal crossing over and mutation is analyzed. This model predicts the spontaneous formation of tandem-repetitive patterns of noncoding DNA from arbitrary sequences for a wide range of parameter values. Natural selection is found to play an essential role in this self-organizing process. Natural selection which is modeled as a mechanism for controlling the length of a nucleotide string but not the sequence itself favors the formation of tandem-repetitive structures. Two measures of sequence heterogeneity, inter-repeat variability and repeat length, are analyzed in detail. For fixed mutation rate, both inter-repeat variability and repeat length are found to increase with decreasing rates of (unequal) crossing over. The results are compared with data on micro-, mini- and satellite DNAs. The properties of minisatellites and satellite DNAs resemble the simulated structures very closely. This suggests that unequal crossing over is a dominant long-range ordering force which keeps these arrays homogeneous even in regions of very low recombination rates, such as at satellite DNA loci. Our analysis also indicates that in regions of low rates of (unequal) crossing over, inter-repeat variability is maintained at a low level at the expense of much larger repeat units (multimeric repeats), which are characteristic of satellite DNA. In contrast, the microsatellite data do not fit the proposed model well, suggesting that unequal crossing over does not act on these very short tandem arrays.

TANDEM-repetitive noncoding DNA sequences make up a large fraction of the genomes of eukaryotes. Sequence complexities can vary from 2 bp to over 2 kb, and the total sizes of tandem arrays range from less than 100 bp to more than 100 Mb (MIKLOS and GILL 1982). According to their repeat lengths and array sizes, tandem-repetitive sequences have been divided into three classes: micro-, mini- and satellite DNAs (LEVINSON and GUTMAN 1987). These different forms of tandem DNA arrays of no obvious functional significance suggest that different DNA turnover mechanisms have been involved in shaping them, while natural selection is usually assumed to play only a minor role.

SMITH (1976) and STEPHAN (1989) have demonstrated by computer simulations that tandem-repetitive structures can be generated *de novo* by the joint action of mutation, unequal crossing over, slippage replication and/or rolling circle replication. Their simulations produced structures for a wide range of parameter values. When recombination rate is varied more than 100-fold, the structures at high rates resemble minisatellites in that the repeats are short and do not show much evidence for higher order structures. In contrast, at low recombination rates, long and more heterogeneous repeats emerge with extensive higher order structures. The computer-generated sequences are qualitatively similar to satellite DNAs, which is consistent with the observation that satellite DNA is concentrated in regions of very low recombination (CHARLESWORTH, LANGLEY and STEPHAN 1986), *i.e.*, in heterochromatic regions near centromeres and telomeres where meiotic recombination is suppressed in many organisms (MATHER 1939). Minisatellite sequences, by contrast, are usually located in the euchromatic portions of chromosomes. Although they are not as uniformly spread across chromosomes as microsatellites but are predominantly found in subtelomeric regions (ROYLE *et al.* 1988), recombination at minisatellite loci appears to be higher than at satellite DNA loci.

Although the earlier study by STEPHAN (1989) found that structural properties of minisatellite and satellite DNAs could be reproduced remarkably well, it was not possible to understand the results of his simulations in terms of standard population genetics theory of multigene families. All population genetic models assume that arrays of tandem repeats pre-exist. The process of self-organization of periodic structures from arbitrary sequences is not considered by these models. Therefore, it was not clear why repeat length and repeat heterogeneity increase when recombination rate decreases (relative to the mutation rate). In particular, the role of selection was not recognized. Selection was built into the model merely as a constraint on sequence length, such that sequences which

grew longer than a certain threshold value were eliminated from the population. But selection did not work on the sequence *per se*. In this paper, we use the theory of self-organization in complex systems (*e.g.*, KAUFFMAN 1993; Chapt. 5) to analyze the properties of the simulation model. This theory helps us to understand the role of selection more clearly and, consequently, also the structural properties of the simulated arrays. To test the predictions of the model, we have compiled relevant data on micro-, mini- and satellite DNAs and analyzed the relationship between the following quantities: repeat length, inter-repeat variability, and allelic heterozygosity due to copy number variation. Our analysis puts the earlier finding that the simulated structures are similar to minisatellite and satellite DNAs on a more quantitative basis. We also find that the simulation results do not fit the microsatellite data, suggesting that unequal crossing over is not a major force in microsatellite evolution.

THE MODEL

We begin by reviewing the model for the generation of tandem-repetitive noncoding DNA proposed by STEPHAN (1989). In the simulation of this model, a single chromosome lineage is followed through time. Three processes are allowed to modify a string of nucleotides according to certain rules during one generation: unequal crossing over by sister-chromatide exchange (SCE), amplification by either slippage replication or a rolling circle mechanism, and mutation by base substitution. The most important rule is that amplification and unequal crossing over depend on sequence similarity; *i.e.*, on local interactions between nucleotide sequences. A minimum length of perfect match between two sequences involved in a DNA turnover process is required for a successful completion of this process. For a rolling circle mechanism sequence similarity is required during the recombinational step leading to the reinsertion of the overreplicated DNA segment into the genome; a recombinational step is obviously also involved in SCE. When the two chromatids mispair in the initial stage of an unequal SCE process, an attempted recombination event can only be successful if the invading strand finds a region of homology on the target strand. In slippage replication sequence similarity is necessary for the reannealing of the daughter strand in an out-of-register position, a process which results in slipped-strand mispairing between the template and daughter strands.

The only amplification process considered here in detail is slippage replication. In modeling slippage as an amplification mechanism, we neglect the propensity of slippage of operating in both directions such that amplification and deletion events may occur. The parameters of the sequence-dependent processes unequal SCE and slippage replication are:

Unequal SCE: m_{rec} = minimum length of a perfect match; γ = rate of SCE for two sequences having a stretch of the minimal match length m_{rec} in common;

Slippage replication: m_{sl} = minimum length of a perfect match for strand reannealing during slippage replication; μ = rate of slippage if the newly synthesized daughter strand and the template have a stretch of m_{sl} nucleotides in common.

In contrast, the mutation process is simple and independent of state; *i.e.*, a nucleotide may change from one state to another according to a time-homogeneous Poisson process with mean u (per generation), and the location of each mutation is randomly distributed along the entire nucleotide string. Four states corresponding to the four bases A, C, G and T are considered. Natural selection works also in a very simple way. The sequence itself is assumed to be functionless, but its length to be controlled by natural selection. Thus, a truncation selection scheme is applied such that the fitness of a sequence is 1, when its length is equal or below a threshold value Ω, and 0 otherwise. 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 generation.

This model is an extension of SMITH's (1976) model. It takes into account the proposition by SMITH that strand exchange in SCE depends on sequence similarity. While SMITH considered only SCE for sequences whose lengths were allowed to vary within a lower bound A ($A > 1$) and an upper one, Ω, our model includes an explicit amplification process. Amplification was introduced to analyze the evolution of a functionless sequence which can be arbitrarily short and, more importantly, to insure an equilibrium length distribution of functionless sequences which is greater than one nucleotide. There are two reasons for this. First, intra- and interspecific studies of simple sequences have demonstrated that tandem-repetitive sequences may persist for a long time; despite no obvious functional significance, they may be present in homologous genome locations in closely related species (*e.g.*, SAVATIER *et al.* 1987). Thus, such sequences may be in a stable or metastable state rather than a transient one, even if they have no function (TACHIDA and IIZUKA 1992). Second, earlier theoretical analyses have shown that the expected life time of simple repetitive sequences under unequal crossing over alone is short because the process gets absorbed quickly at copy number one in finite populations (STEPHAN 1986; WALSH 1987). Thus, it appears that forces exist which lead to sequence amplification at least occasionally.

The introduction of a sequence-dependent amplification process had important consequences. STEPHAN (1989) demonstrated by computer simulations that the proposed model of slippage replication and unequal SCE leads to spontaneous formation of tandem-repetitive patterns for a wide range of recombination rates γ, while a model of unequal SCE alone predicts periodicities only for relatively high recombination values. This is due to the synchronization of these two mechanisms which are coupled via the sequence similarity rule. It was concluded that this model may explain the existence of periodic structures of no obvious functional significance, such as minisatellites, which are usually located in regions of intermediate to high recombination rates in euchromatin. Most interestingly, this model may also explain the periodic structure of satellite DNAs which are thought to exist only in regions of very low recombination (heterochromatin). For satellite DNA, rolling circle replication may be the dominant amplification mechanism. Extrachromosomal circular satellite DNA has been found in some instances (OKUMURA, KIYAMA and OISHI 1987), suggesting that rolling circle amplification of satellite sequences occurs. On the other hand, the length of slipped-strand mispairing appears to be restricted (LEVINSON and GUTMAN 1987). It may therefore be more appropriate to replace (or perhaps extend) the slippage process by a rolling circle mechanism. However, due to the sequence similarity assumption, the results based on a biased slippage mechanism can be expected to be qualitatively the same as those of a rolling circle model.

EVOLUTION OF TANDEM-REPETITIVE DNA AS A SELF-ORGANIZING PROCESS

The earlier simulations have demonstrated that the formation of periodic patterns occurs spontaneously and appears to be independent of the initial state of the nucleotide string. STEPHAN (1989) did most of the simulation runs starting from a string of identical nucleotides. However, runs started from a random sequence led also to repetitive structures and overall to similar results for a given parameter set. SMITH (1976) started his simulations with random sequences. Thus, the system can move toward a periodic structure starting from a very ordered initial state where the nucleotide string is homopolymeric, as well as from a chaotic initial state in which the sequence is random. This phenomenon seems to characterize complex systems that tend toward self-organization. For instance, KAUFFMAN demonstrated such behavior using random Boolean networks (reviewed in KAUFFMAN 1993; Chapt. 5). Complex networks are defined by local features that describe the interactions between elements within a network (analogous to our sequence similarity rule) and by a scoring system. The scoring system is important in that it mimics natural selection by determining which networks "reproduce" preferentially in comparison with others. In our model, natural selection which enters in a seemingly simple way as a constraint for sequence length plays a pivotal role in that it drives the system away from the chaotic region of the state space and also away from a completely ordered state toward an intermediate region which exhibits a relatively high degree of order. This subtle action of natural selection which was not recognized in the previous study (STEPHAN 1989) can be understood as follows. As outlined above, selection does not work on sequence *per se*. It is thus not important what the sequence of the nucleotide string is. All natural selection does is to favor shorter sequences. When a sequence grows longer than Ω by unequal SCE or amplification, it is eliminated. In the simulation of a chromosome lineage, this is realized by choosing the sister strand (whose length is shorter than Ω) for continuing the simulation. Thus, shorter sequences are passed on preferentially to the next generation. This ties in with the evolution of repetitive structures. On average, sequences which evolve a repetitive structure have a higher chance to undergo unequal exchange because of the similarity requirement. More frequent crossing overs keep sequences farer away from the boundary Ω. This can be shown analytically (STEPHAN 1987). Lineages which develop periodicities can therefore survive better than lineages with nonrepetitive sequences. The rather general conclusion from these models by SMITH (1976) and STEPHAN (1989) is that natural selection can push random sequences toward order by controlling sequence length. Sequences which evolve a repetitive pattern can recombine more often and thus avoid elimination.

It is also possible to understand why our model leads to spontaneous formation of periodic patterns independent of the initial state of the sequence. If the sequence is completely ordered such that it consists of identical nucleotides, then misalignment of strands during slippage replication and/or unequal SCE always satisfies the similarity requirement. A pattern can begin to emerge when a mutation has occurred at some site in the interior of the sequence because homopolymeric stretches of nucleotides still exist and thus there is opportunity for strand attachment. This situation is depicted in Figure 1 for unequal exchange between sister chromatids. Given that occasional mutations alter the local characteristics of a sequence, such homopolymeric stretches can also be created when the initial sequence was random. Then, on a local scale, pattern formation can occur in the same way as described for the completely ordered sequence. The emerging local pattern may function as a nucleation site for further growth. The growing phase appears to be essentially determined by unequal crossing

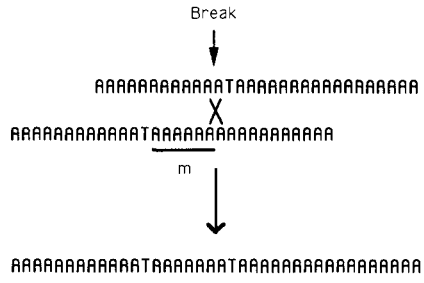

FIGURE 1.—Formation of a TAAAAAAA repeat unit by unequal sister strand exchange. The chromatids are paired out of register. A break creates a free end which can find a stretch of m nucleotides of perfect match for strand annealing. As shown, this results in a daughter chromatid with two TAAAAAAA repeats.

over which can involve large numbers of repeats and hence act as a long-range ordering force, while slippage seems to be more restricted and to work only over shorter distances. This is in agreement with an hypothesis which was advanced by LEVINSON and GUTMAN (1987) based on a data set of tandem-repetitive DNA ranging from microsatellites to satellite DNA.

The theory of self-organization in complex systems can also be used to understand in a coherent way several specific properties of the simulated structures which can only partially be explained on the basis of standard population genetic theory of multigene families. These include: (i) the tendency to increased inter-repeat variability and (ii) longer repeat units with decreasing recombination rate γ and (iii) the tendency to longer repeat units and increased inter-repeat variability for increasing Ω. These properties were noticed in the original simulations (STEPHAN 1989) and were confirmed in this study. A summary of the simulation results is displayed in Figure 2.

The third property can immediately be understood from our discussion of the role natural selection given above. With decreasing Ω and thus an increasing strength of selection, the structures that survive are those which allow unequal SCE more often, *i.e.*, which have shorter and more homogeneous repeats. Thus the upper limit set by selection allows the system to evolve away from the chaotic region of the sequence space into a somewhat ordered regime. This may have biological implications in that some genome regions (*e.g.*, euchromatin) experience stronger selection (*i.e.*, a smaller upper limit) than others (*e.g.*, heterochromatin), suggesting that the structural differences between euchromatic and heterochromatic, tandem-repetitive sequences are at least partially due to selection (see below).

An interesting corollary of the effect of selection is the formation of subarrays within the total nucleotide string. This is illustrated in Table 4 of STEPHAN (1989). For that parameter set, 4 out of 10 simulations led to 2 distinctly different subarrays. Since this resulted in a reduced effective Ω for each of these subarrays, inter-repeat variability in the subarray-containing structures should be lower than in the unpartitioned ones, which was indeed observed. We repeated these simulations for the same parameter set and found an inter-repeat variability of 0.8 ± 0.5% in the subarray-containing nucleotide strings *vs.* 2.8 ± 1.1% in the strings with no subarrays.

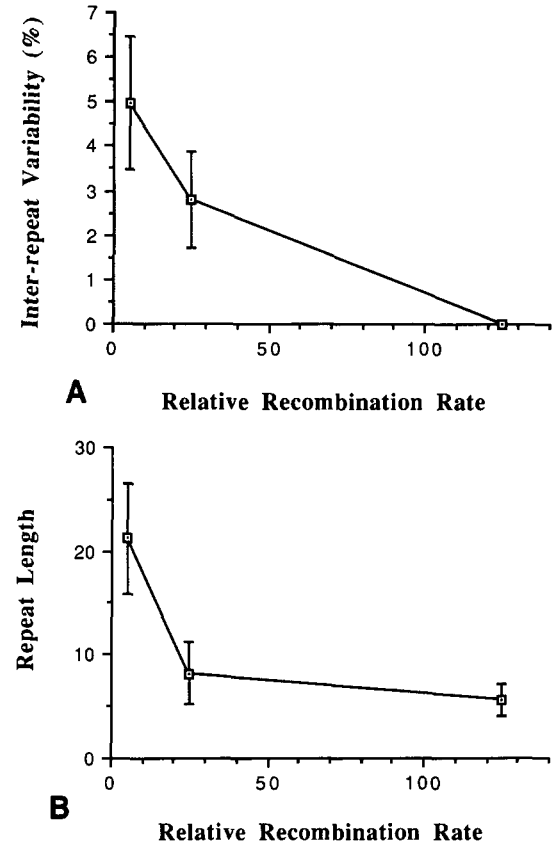

FIGURE 2.—Sequence heterogeneity *vs.* relative recombination rate γ/u. (A) Inter-repeat variability decreases with increasing rates of (unequal) crossing over. (B) Repeat length (in bp) decreases with the rate of (unequal) crossing over. The differences in the slopes for small and large recombination rates indicates that both measures of sequence heterogeneity are influenced by changes in γ/u in different ways (see text). The parameter values are as follows: u = 0.00005; $\mu/\gamma = 100$; $m_{rec} = m_{xl} = 5$; $\Omega = 1000$.

The first two properties of the model can be understood by quantifying sequences with regard to randomness/order. For a given sequence length, it is clear what the extreme ends of the sequence space are with regard to randomness. The homopolymeric sequences are the states of complete order, random sequences make up the chaotic region. Distance of any state in the sequence space to these extremes may be measured by repeat length, such that the order of a sequence decreases with repeat length. If the repeat length is one nucleotide, the sequence is homopolymeric, and if it is greater than the sequence length, the sequence is random. If a repetitive structure is observed, a reasonable metric would be to measure

randomness in terms of inter-repeat variability. Using these definitions, it becomes clear what the simulation results in Figure 2 mean. As the relative recombination rate, *i.e.*, the ratio γ/u, decreases the strength of the ordering force of unequal crossing over becomes weaker relative to the randomizing mutation process. This leads to more random repetitive structures, *i.e.*, longer repeat units and larger inter-repeat variability, as seen in the simulations.

The simulations summarized in Figure 2 reveal more interesting details. The differences in the slopes of the curves in Figure 2 for different values of γ/u suggest that both measures of randomness, repeat length and inter-repeat variability, are not equivalent. Between $\gamma/u = 25$ and 125, inter-repeat variability appears to be a more sensitive indicator of differences in γ/u. In contrast, for low γ/u, repeat length decreases more rapidly than inter-repeat variability. The following is going on. For large γ/u, the system reacts to the randomizing forces of mutation by producing periodicities, while inter-repeat variability is kept at a very low level. The level of inter-repeat variability changes more quickly as γ/u is lowered, whereas average repeat length does not change much. Further lowering of γ/u leads to an interesting second phase for $\gamma/u < 25$, where repeat length increases rapidly, while inter-repeat variability changes only slowly. In this second phase, higher order periodicities (multimers) composed of basic repeat units (monomers) are formed. Thus, we see a second level of self-organization in which higher order structures are created: Inter-repeat variability (between the multimeric units) is kept at a relatively low level at the expense of larger repeat units (multimers). The reason for this phenomenon appears to be clear. Sequences which preserve a repetitive structure are still able to undergo SCE and slippage, even when the repeats are longer. This would not be possible if the system would react to increased mutation pressure differently by producing more heterogeneous repeats (instead of longer ones). It appears that the system tunes itself to the point where it can stay away from the (chaotic) subspace of random sequences. The higher order phenomena, such as the formation of multiple subarrays within the nucleotide string (discussed above) and the higher order structures (multimers), are examples for the "creativity" of the system in this region of the phase space.

THE DATA

Next we examine whether the properties of this model can be seen in the available data. We focus on the human data set which is the most comprehensive one. It contains tandem-repetitive noncoding DNAs of very short (microsatellite DNA) to long repeat units (α satellite DNA). Many properties of the simulated structures resemble those of minisatellites and satellite DNAs, whereas our model appears to be less important for the microsatellite data.

There is experimental evidence that the proposed mechanisms of SCE, slippage replication and rolling circle replication work on minisatellite and satellite DNA clusters. Extrachromosomal circular satellite DNA has been observed (OKUMURA, KIYAMA and OISHI 1987), suggesting that rolling circle replication may be an important amplification mechanism for satellite DNA. For minisatellites, slippage appears to be the predominant mode of new allele formation (WOLFF *et al.* 1989), however evidence for unequal sister chromatid exchange has also been found in the data (JEFFREYS *et al.* 1991). The latter two mechanisms have been modeled here such that the intrahelical process of slipped-strand mispairing is much more likely to act on short nucleotide strings and that, after the initial expansion of the sequence beyond a certain threshold, unequal crossing over can also work on the array. Unequal SCE is an interhelical event, involving DNA molecules from two different sister chromatids. This may place some constraints on the length of the sequences undergoing unequal SCE (LEVINSON and GUTMAN 1987). The microsatellite data suggest that repetitive tracts have to be longer than 100 bp for unequal SCE to work (see below).

Minisatellite DNAs: Figure 3 shows a compilation of human minisatellite data. We included variable minisatellite loci (VNTRs) whose repeat lengths were at least 15 bases, and for which sufficient information was available on allelic heterozygosity due to copy number variation and on sequence heterogeneity within an array. The criterion of 15 bp is to some extent arbitrary. There are tandem-repetitive arrays with repeat lengths <10 bp (*e.g.*, locus *D1S7*; JEFFREYS *et al.* 1988) which are hypervariable and have a large copy number. Sometimes such loci are classified as minisatellites. Most classifications, however, are closer to ours. Choosing only tandem-repetitive loci with this larger repeat length seems to increase the relative importance of unequal SCE over slippage replication, because there appears to be a constraint on the length of slipped-strand mispairing during replication (LEVINSON and GUTMAN 1987).

Based on these data the relationship between the following quantities was examined: repeat length, inter-repeat variability, and allelic heterozygosity due to copy number variation. Allelic heterozygosity of the VNTRs included in the analysis ranges from 30 to 99%. Most well studied minisatellite loci are used for DNA typing. Thus, Figure 3 is not a representative sample of minisatellites in humans. There is a bias toward high variability loci. The most striking result of our analysis is that there is a significantly negative correlation between inter-repeat variability and pop-

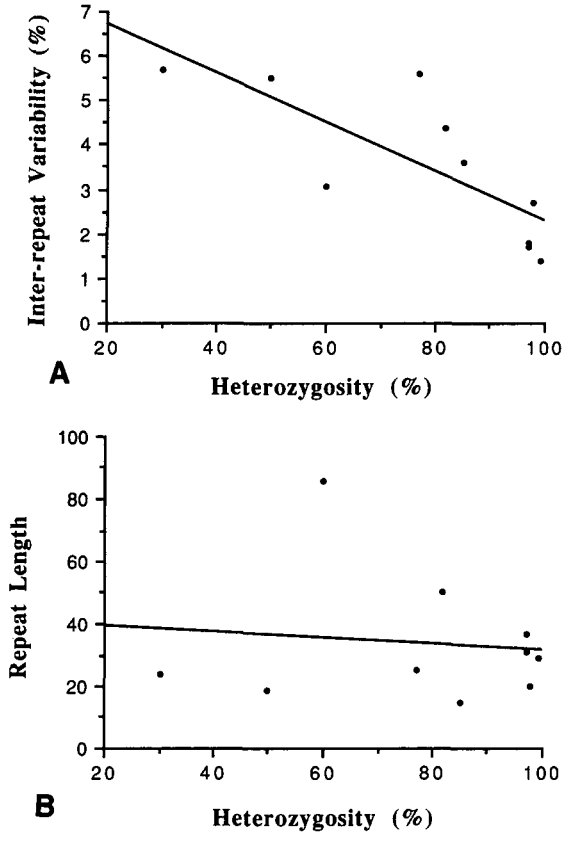

FIGURE 3.—Sequence heterogeneity *vs.* copy number variation at human minisatellite loci. (A) The correlation between inter-repeat variability and allelic heterozygosity is significant (R^2 = 0.602; $P < 0.01$). (B) The correlation between repeat length and population heterozygosity is not significant (R^2 = 0.012; NS). The data bases GenBank (release 75.0) and EMBL (release 33.0) were searched with the program TURBOGOPHER (University of Minnesota, Computer and Information Services) for VNTRs. The keywords "minisat," "mini-sat" and "VNTR" were used in this search. Minisatellite loci were included in the graphs if (i) repeat length was ≥15 bp, (ii) a sufficient number of sequences (at least five repeats) were found in the data bases for estimating inter-repeat variability and (iii) estimates of allelic heterozygosity were available from large samples of individuals (usually more than 50). The data on allelic heterozygosity are from SILVA, JOHNSON and WHITE (1987), JEFFREYS *et al.* (1988), ARMOUR *et al.* (1989), INGLEHEARN and COOKE (1990), JEFFREYS *et al.* (1991), and SCOTT *et al.* (1991).

ulation heterozygosity (Figure 3A). Thus, loci showing the greatest level of allelic variation in copy number show low levels of inter-repeat variability, presumably resulting from sequence homogenization of repeat units by unequal crossing over. In contrast, there is no correlation between repeat length and allelic heterozygosity (Figure 3B). Both observations appear to be in accordance with the predictions of the model (Figure 2). Higher copy number variation suggests more frequent unequal exchanges. Since inter-repeat variability is not affected, higher levels of heterozygosity are presumably due to higher values of γ/u. In the range of intermediate to large values of γ/u, the model predicts that inter-repeat variability depends more strongly on changes of γ/u than repeat length (compare Figure 2, A with B).

Satellite DNAs: For satellite DNA, there are no good data available on copy number variation within populations. Nonetheless, it is interesting to compare the properties of satellite DNA with those of minisatellites. When recombination rate γ is significantly reduced, both repeat length and inter-repeat variability should be increased. A comparison between Figure 2, A and B, suggests that the effect is more dramatic for repeat length than inter-repeat variability, contrary to the behavior for large and intermediate values of γ/u. Thus, for satellite DNA clusters, we should expect that repeat units are much longer than for minisatellites, whereas inter-repeat variability is only slightly increased.

It appears that the condition of low recombination rates is generally met for satellite DNAs. Satellite DNA forms large clusters of tandem-repetitive structures which are usually located in centromeres and telomeres. There is good evidence from Drosophila that recombination rate is severely repressed near centromeres and often also near telomeres (MATHER 1939; LINDSLEY and SANDLER 1977). The evidence for other species is less clear. However, WEVRICK and WILLARD (1989) report that SCE during mitosis or meiotic recombination in human alphoid DNA regions is also low. In humans, the α satellite DNA family is best studied. It is organized as long arrays (up to 1–3 Mb) of tandem-repetitive units in the centromeres of all chromosomes (WILLARD 1985). The chromosome subfamilies display a higher order structure based on multimers of the 171 bp alphoid monomer repeat unit. A multimeric repeat is of the form (*ABC* . . .), where *A*, *B*, *C* are variants of a basic monomer which differ distinctly from each other. The linear multimeric arrays of human α satellite DNA generally show substantial inter-monomer sequence variability of the order 20–40%. However, corresponding monomers, say the variants *A*, between pairs of multimeric repeat units show considerably less variability. For instance, the human *X* chromosome α repeat has a higher order structure consisting of a repeated array of 12 distinct 171-bp monomers which differ on average by 24%. However, corresponding monomers between two 12 mers differ only by about 4% (LAURSEN *et al.* 1992). The latter value for inter-repeat variability in human α repeats is relatively low and not much larger than the average value of the hypervariable minisatellite loci (see Figure 3A). The values of inter-repeat variability of the major satellites from several species (including Drosophila species) compiled by MIKLOS and GILL (1982) are similar. Recently BACHMANN and SPERLICH (1993) reported inter-repeat variability values of 4–6% for a large number of randomly cloned repeats of a monomeric

182-bp satellite DNA family from the centromeric heterochromatin of three species of the *Drosophila obscura* group. Higher values of inter-repeat variability (of more than 10%) have been estimated in other satellite families of the *obscura* group, in particular, in longer tandem arrays (L. BACHMANN, personal communication). Such a positive correlation between average copy number of an array and inter-repeat variability would be in accordance with our model for the following reason. Assuming that the upper limit to copy number, Ω, is approximately equal for all satellite families, less frequent unequal exchanges due to higher levels of inter-repeat variability should lead to an accumulation of tandem-repetitive DNA (STEPHAN 1986, 1987). Obviously, this correlation could also be explained if array size scales with Ω.

In contrast to inter-repeat variability, repeat lengths of satellite DNAs are mucher larger than those of minisatellites. In invertebrates repeats as long as long as 359 bp have been found, in vertebrates the longest known repeat is 2350 bp (reviewed in MIKLOS and GILL 1982). The very long repeats, such as the latter ones, are usually composed of subunits. In general, repeats range between 100 and 400 bp. Although the satellite DNA data set is not as comprehensive as that for minisatellite loci, it appears that satellite repeats are on average much longer but only slightly more heterogeneous than those of the hypervariable minisatellite loci. This observation is in agreement with the proposed model. As described above, this model predicts that inter-repeat variability and repeat length are influenced in different ways for small values of γ/u, such that multimeric repeats arise and repeat length becomes much longer, while sequence heterogeneity stays almost constant (see Figure 2). Formation of higher order repeats, such as dimers, trimers and longer multimers, is generally seen in satellite DNA clusters, but less frequently in minisatellites.

The similarity between the simulated structures and the data goes even further and includes subtle structural details. A feature of the simulated repeats is the occurrence of homopolymeric nucleotide tracts within repeat units. This is also seen in the data, in both minisatellite and satellite DNAs. It is likely that these stretches of identical nucleotides are reminescent of the initial phase of the *de novo* synthesis of repetitive patterns. As shown in Figure 1 and discussed above, patterns are most easily created by slipped-strand mispairing on homopolymeric nucleotide stretches. This feature is discussed in detail in the previous paper (STEPHAN 1989).

Microsatellite DNAs: Microsatellite arrays consist of runs of very short repeat units (of less than 10 nucleotides). The total length of a run is usually less than 100 bp. Often there are several such runs found at a single microsatellite locus. Table 3 of RIGGINS *et al.* (1992) gives instructive examples of microsatellite loci (CGG repeats) with flanking regions. From these examples it appears that perfect runs of CGG repeats are shorter than 50 bp. After being interrupted, the same repeats are often continued. The intervening sequence may be random or itself a short run of a 3-bp repeat of a different motif. This suggests that the evolution of these simple sequences is not undergoing long-range ordering forces such as unequal crossing over. Given that microsatellites are distributed across the whole chromosomes (including regions of intermediate to high crossing over), unequal exchanges should be a very powerful ordering force, if they acted on these sequences. However, without the action of unequal crossing over, there is only slippage replication which works on such sequences. Slippage alone is not a strong force of pattern formation. As outlined above, it is the synchronization of sequence-dependent amplification and unequal crossing over which leads to efficient homogenization.

There is direct evidence that slippage works on microsatellites (SCHLÖTTERER and TAUTZ 1992). A straightforward prediction of the slippage model is that this process occurs more frequently in longer arrays. The best way to examine this prediction would be to test whether there is a significant correlation between the total length of a run of perfect repeats and copy number variation in a population. Indeed, for eight human CGG repeat 5′-UTR loci a correlation coefficient of 0.78 was determined comparing the uninterrupted repeat length of the most common allele with the average polymorphic information content, a measure of heterozygosity (RIGGINS *et al.* 1992). Among these eight loci was the CGG-repeat in the 5′-UTR of the *FMR-1* locus which is unstable in fragile *X* families. In contrast, VALDES, SLATKIN and FREIMER (1993) did not find a correlation between the variance in allele size at a locus and the mean allele size for dinucleotide repeat loci. This may be partly due to the fact that the latter authors did not use the repeat numbers of uninterrupted runs in their analysis but the total lengths of microsatellite loci (including flanking regions).

Evidence for selection: The effect of natural selection on simple sequence DNA can be expected to vary across chromosomes in that the upper limit to sequence length, Ω, set by selection may be smaller in some genome regions than others. Our model predicts that stronger selection (*i.e.*, a smaller upper limit) has a similar effect on sequence structure as increased recombination rates. This may make it difficult to distinguish the effects of unequal crossing over from those of selection. For instance, the structural differences between euchromatic and heterochromatic tandem-repetitive sequences (*e.g.*, between human minisatellite and satellite DNA sequences) may be due to

both stronger selection and increased recombination rates in euchromatic regions. However, the two effects may be distinguished in species, such as *Drosophila melanogaster*, in which recombination rate varies greatly not only between heterochromatin and euchromatin, but also within the euchromatic portions of the chromosomes (LINDSLEY and SANDLER 1977). In *D. melanogaster* euchromatin, long tandem arrays such as minisatellites have not been found, although recombination rate is reduced over large portions of the genome. In contrast, many highly repetitive DNA sequences have been mapped in heterochromatin of *D. melanogaster* (LOHE, HILLIKER and ROBERTS 1993), suggesting that the upper limits to sequence length vary greatly between heterochromatin and euchromatin in *D. melanogaster*. This indicates that the distinct structural differences between euchromatic and heterochromatic tandem arrays observed in most eukaryotes are at least partly due to selection.

DISCUSSION

Two properties of the simulated tandem-repetitive structures were explored in this paper: the tendency to longer repeat units and larger sequence heterogeneity between repeats with decreasing recombination rate. It was concluded that longer repeat units and larger inter-repeat variability indicate a higher degree of randomness of the sequence and that lower recombination rates (relative to the mutation rate) increase randomness. Higher order structures, such as multimeric repeats and multiple subarrays, characterize the system's response at lower recombination rates. We compared these properties with tandem-repetitive noncoding DNAs ranging from micro- and minisatellites to satellite DNAs. While micro- and minisatellites are thought to be located in euchromatic regions of intermediate to high recombination rates, satellite DNAs are usually found in the heterochromatin in which recombination is severely suppressed. The properties of minisatellite and satellite DNAs which form long clusters compare well with the features of the simulated structures. In accordance with the model, we found that repeat length in satellite DNAs is on average much longer than in minisatellites and that higher order structures are common in satellite DNAs but not in minisatellites. In contrast, inter-repeat variability in satellite arrays is only slightly higher than in minisatellite loci. However, the proposed model of sequence-dependent amplification and unequal SCE does not fit well the microsatellite data. While it is likely that slippage replication is working on microsatellite loci, the array sizes of these loci appear to be too short for unequal crossing over which involves two separate molecules and may thus be sterically more constrained than slippage (LEVINSON and GUTMAN 1987). For microsatellites, birth-death models appear to be sufficient (WALSH 1987; TACHIDA and IIZUKA 1992; VALDES, SLATKIN and FREIMER 1993).

We interpreted the results of a simulation model for the spontaneous formation of repetitive DNA (STEPHAN 1989) in the framework of complex dynamical systems that are able to evolve. We found that selection introduced in the model as a constraint on sequence length favors repetitive structures. Those sequences in which repetitive structures emerge have a higher chance to survive the selection process. This is an interesting shift in the role of natural selection from selection against a certain trait (sequence length) of a replicating unit to *apparent* selection on another trait which is not causally connected to fitness (repetitive structure). It raises many questions. For instance, if self-organization played a major role in evolution, selection may work in more intricate ways than previously thought. A certain trait which has evolved and is thought to be an adaptation, may not have been molded by selection in a direct way. Instead, selection may have worked on something different, but indirectly favored this trait. If so, this may have far-reaching consequences for the theory of adaptation. At present, however, we are only beginning to understand the evolutionary potential of self-organization in conjunction with natural selection.

We thank LUTZ BACHMANN (Universität Tübingen) for sharing his unpublished satellite DNA sequence data, JAQUELINE HARTZLER and SEAN RYDER for assistance in compiling the minisatellite data, and the 1992 Molecular Evolution class (ZOOL 441/645) for insightful discussions.

LITERATURE CITED

ARMOUR, J. A. L., Z. WONG, V. WILSON, N. J. ROYLE and A. J. JEFFREYS, 1989 Sequences flanking the repeat arrays of human minisatellites: association with tandem and dispersed repeat elements. Nucleic Acids Res. **17:** 4925–4935.

BACHMANN, L., and D. SPERLICH, 1993 Gradual evolution of a specific satellite DNA family in *Drosophila ambigua, D. tristis,* and *D. obscura*. Mol. Biol. Evol. **10:** 647–659.

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.

INGLEHEARN, C. F., and H. J. COOKE, 1990 A VNTR immediately adjacent to the human pseudoautosomal telomere. Nucleic Acids Res. **18:** 471–476.

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., A. MACLEOD, K. TAMAKI, D. L. NEIL and D. G. MONCKTON, 1991 Minisatellite repeat coding as a digital approach to DNA typing. Nature **354:** 204–209.

KAUFFMAN, S. A., 1993 *The Origins of Order–Self-Organization and Selection in Evolution*. Oxford University Press, New York.

LAURSEN, H. B., A. L. JØRGENSEN, C. JONES and A. L. BAK, 1992 Higher rate of evolution of X chromosome α-repeat DNA in human than in the great apes. EMBO J. **11:** 2367–2372.

LEVINSON, G., and G. A. GUTMAN, 1987 Slipped-strand mispair-

ing: A major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4:** 203–221.

LINDSLEY, D. L., and L. SANDLER, 1977 The genetic analysis of meiosis in female *Drosophila melanogaster*. Philos. Trans. R. Soc. Lond. **277B:** 295–312.

LOHE, A. R., A. J. HILLIKER and P. A. ROBERTS, 1993 Mapping simple repeated DNA sequences in heterochromatin of *Drosophila melanogaster*. Genetics **134:** 1149–1174.

MATHER, K., 1939 Crossing over and heterochromatin in chromosomes of *Drosophila melanogaster*. Genetics **24:** 413–435.

MIKLOS, G. L. G., and A. C. GILL, 1982 Nucleotide sequences of highly repeated DNAs; compilation and comments. Genet. Res. **39:** 1–30.

OKUMURA, K., R. KIYAMA and M. OISHI, 1987 Sequence analyses of extrachromosomal *Sau*3A and related family DNA: analysis of recombination in the excision event. Nucleic Acids Res. **15:** 7477–7489.

RIGGINS, G. J., L. K. LOKEY, J. L. CHASTAIN, H. A. LEINER, S. L. SHERMAN, K. D. WILKINSON and S. T. WARREN, 1992 Human genes containing polymorphic trinucleotide repeats. Nature Genet. **2:** 186–191.

ROYLE, N. J., R. E. CLARKSON, Z. WONG and A. J. JEFFREYS, 1988 Clustering of hypervariable minisatellites in the proterminal regions of human autosomes. Genomics **3:** 352–360.

SAVATIER, P., G. TRABUCHET, Y. CHEBLOUNE, C. FAURE, G. VERDIER and V. M. NIGON, 1987 Nucleotide sequence of the β-globin genes in gorilla and macaque: the origin of nucleotide polymorphisms in human. J. Mol. Evol. **24:** 309–318.

SCHLÖTTERER, C., and D. TAUTZ, 1992 Slippage synthesis of simple sequence DNA. Nucleic Acids Res. **20:** 211–215.

SCOTT, H. S., P. V. NELSON, J. J. HOPWOOD and C. P. MORRIS, 1991 PCR of a VNTR linked to mucopolysaccharidosis type I and Huntington disease. Nucleic Acids Res. **19:** 6348.

SILVA, A. J., J. P. JOHNSON and R. L. WHITE, 1987 Characterization of a highly polymorphic region 5′ to J_H in the human immunoglobulin heavy chain. Nucleic Acids Res. **15:** 3845–3857.

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.

STEPHAN, W., 1987 Quantative variation and chromosomal location of satellite DNAs. Genet. Res. **50:** 41–52.

STEPHAN, W., 1989 Tandem-repetitive noncoding DNA: forms and forces. Mol. Biol. Evol. **6:** 198–212.

TACHIDA, H., and M. IIZUKA, 1992 Persistence of repeated sequences that evolve by replication slippage. Genetics **131:** 479–491.

VALDES, A. M., M. SLATKIN and N. B. FREIMER, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133:** 737–749.

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

WEVRICK, R., and H. F. WILLARD, 1989 Long-range organization of tandem arrays of α satellite DNA at the centromeres of human chromosomes: high-frequency array-length polymorphism and meiotic stability. Proc. Natl. Acad. Sci. USA **86:** 9394–9398.

WILLARD, H. F., 1985 Chromosome-specific organization of human α satellite DNA. Am J. Hum. Genet. **37:** 524–532.

WOLFF, R. K., R. PLAETKE, A. J. JEFFREYS and R. WHITE, 1989 Unequal crossing over between homologous chromosomes is not the major mechanism involved in the generation of new alleles at VNTR loci. Genomics **5:** 382–384.

Communicating editor: M. BULMER