Neutral Mutations and Repetitive DNA

William F. Loomis and Michael E. Gilpin

Received July 30, 1987

KEY WORDS: computer simulation; neutral mutations; repetitive DNA.

We have previously shown that computer simulations of processes that generate selectively advantageous changes together with random duplications and deletions give rise to genomes with many different genes embedded in a large amount of dispensable DNA sequence. We now explore the consequences of neutral changes on the evolution of genomes. We follow the consequences of sequence divergences that are neutral when they occur in dispensable sequences or extra copies of genes present in multigene families. We find that when divergence occurs at about the same frequency as duplication/deletion events, genomes carry repetitive sequences in proportion to their size. Inspection of the genomes as they evolved showed that multigene families were generated by relatively recent duplications of single genes and so would be expected to be highly homogeneous.

INTRODUCTION

Repetitive DNA sequences have been found in the genomes of all eukaryotic organisms (1, 2, 3). When double-stranded DNA of a given species is melted to its composite single strands and then allowed to reanneal under defined conditions, a considerable proportion reforms double-stranded DNA more rapidly than would be expected for sequences present only once in the genome. Although much speculation has centered on possible functions for repetitive DNA, it is still a component without a physiological role (1, 2, 4, 5). We suggest that most repetitive DNA is a by-product of evolving genomes with no selective value and that it is a negligible price to pay to allow rare but selectively advantageous changes in duplicated copies of genes.

There appears to be a continuum of sequence relationships among members of the repetitive class with some sequences having almost exact copies and other sequences having only distantly related copies (6). Highly homologous sequences reanneal as repetitive DNA even under the most stringent conditions and the duplexes that are formed...
formed melt only at the temperature (Tm) observed for genomic DNA. Other sequences reanneal as repetitive DNA under less stringent conditions and have a lower Tm. These sequences are clearly related but have accumulated base changes since the time they duplicated. Some organisms have only 3% repetitive DNA while others have up to 70% repetitive DNA (1, 7). The proportion appears to depend on the size of the genome.

The rate of base change in vertebrates has been determined in untranslated regions and in pseudogenes to be 5 ± 3 x 10^-9 per base per year (8). In almost all cases these would be expected to be neutral changes. Therefore, a genome of 10^9 bases sustains several base changes per year. This rate is considerably higher than the rate of duplication/deletion events which we have estimated to occur once every 10^3 years in such a genome (9). Duplications generate repeat units that are initially exact copies of the original units while deletions remove units that often have drifted from their original units. Dispensable units accumulate fine-scale rearrangements and mutations continuously, and after a few hundred base changes have occurred in a unit of 1 kilobase, they will no longer reanneal with copies of the original unit under the standard conditions of stringency; that is, they will no longer be members of the repetitive class.

Duplicated units can drift considerably from the original sequence and yet reanneal under low stringency conditions. At some point they will become susceptible to changes that preclude reannealing to other members of their repetitive family. Susceptible members of the family can be recognized as that portion of the repetitive DNA which is only observed at the least stringent conditions for reannealing. The results of reannealing studies on complex genomes do not have sufficient resolution to determine accurately the proportion of repetitive DNA that is highly susceptible to divergence, and it is not easy to calculate the proportion from the measured rates of base changes and duplication/deletion events because of their stochastic nature. Therefore, we define a divergence event as the change that empirically results in a unit leaving a repetitive family and analyze genomes in which the ratio of divergence to duplication/deletion events is varied. Divergence events are only permitted in the dispensable DNA, which includes vestigial sequences and extra copies of genes. We are not concerned here with localized simple sequence, highly repetitive DNA that is generated by run-away unequal crossing over (10, 11). Nor are we concerned with retroviruses, transposons or other examples of selfish DNA that have specific mechanisms of replication (12, 13).

We have determined the consequences of varying the rate of sequence divergence to the rate of duplication/deletion events in both large and small genomes. We find that, by allowing dispensable sequences generated by our model to diverge, repetitive DNA can be accounted for by defined parameters and that the proportion of repetitive DNA is greater in large genomes than in small ones.

**METHODS: The GENESIS Program**

A program was written in True BASIC following the same parameters that generated complex genomes in our previous study (9). The program is compatible with
either Macintosh or IBM compatible personal computers and is available on disc upon request from the authors. A simulation of 10,000 events can be completed in half an hour and the data stored for further analysis. The program represents a gene by 4 contiguous symbols that must start with \#, representing a promoter, and must end with a period representing a termination signal. Between these symbols are letters that designate different genes. Each symbol can be thought of as a kilobase of DNA. Complex genomes are generated by Selectively Advantages Changes (SAC) that convert the second letter in a gene to the next in the alphabet, for example, \# AA becomes \# AB. Only genes present in two or more copies are permitted to undergo a SAC so as to leave at least one original gene copy in the genome. Genomes with the greatest number of different genes are followed. The program can start with a single gene such as \# AA or with a complex genome. To avoid skewing our simulations in this study we started with 52 contiguous genes represented by two identical letters in upper or lower case. The program chooses a position in the genome at random and picks an arbitrary number of units to either duplicate or delete. Unless otherwise mentioned, the probability of duplication was set equal to that of deletion. The maximum number of units affected was set at 20. Deletions were only accepted if, after the event, the genome retained at least one copy of all previously existent genes. Duplications could occur anywhere and were positioned tandemly.

Duplications produced redundant copies of genes that were available for a SAC. We usually set the SAC rate at 0.1/G at each duplication/deletion event where G was the number of genes present at that time in the genome. Every 10 events on average the program would attempt to carry out a SAC and would do so if the gene that was chosen at random was redundant. In this way the genome became increasingly complex, simulating the evolution of a genome; however, we found that reducing the complexity had little effect on the proportion of repetitive sequences in the genome. The program generates genomes of about 13,000 units containing more than 100 different genes after 10,000 events. Each gene is present in about 1.5 copies on average and 95% of the genome consists of dispensable vestigial sequences that were generated by duplication of incomplete genes and deletions of portions of redundant genes (9). So as to analyze families of repetitive sequences, divergence of duplicated units was simulated in the GENESIS program. A prefix was added to each unit to distinguish the original functional units from those that had drifted. Sequence divergence (SD) changed the prefix of each letter to a higher number. For instance, a unit (1A) that was chosen for SD became 2A. If that unit was chosen again, it became 3A and so on up to 24 different prefixes. Only letter units with the same prefixes were considered members of the same class. GENESIS can generate 1392 different classes although less than 1000 appeared even in simulations run for 100,000 events. When a unit at the highest state is chosen for another SD event, it is eliminated from further consideration. A unit can sustain a SD event if it is present as a vestigial sequence or is part of a redundant gene. Therefore, all SD events are neutral to the genetic complexity of the genome.

The SD rate was set relative to the duplication/deletion rate so as to generate genomes with varying proportions of repetitive sequences. The distribution of classes with different numbers of repetitive sequences were compared in large and small genomes. Genome size was regulated by the ratio of duplications to deletions. When
Fig. 1. Distribution of repetitive sequences. Genomes were generated by 20,000 events at various SD settings with SAC set at 0.1/G. The number of classes present at the same copy number was determined.
deletions were allowed to occur more frequently than duplications, genomes remained small and had fewer repetitive sequences.

RESULTS AND DISCUSSION

Sequence Divergence

The distribution of sequences present in up to 100 copies was found to be significantly affected by the rate of sequence divergence (Fig. 1). When no divergence occurred, most classes were present at high copy number after 10,000 events and the distribution did not change significantly after another 10,000 events. Only 1% of the complexity was present in three copies or less and 90% was present in 10 copies or more after 20,000 events. A third of the sequences were present in greater than 100 copies each. When SD was set at 0.5, the number of classes increased and the mean copy number decreased (Fig. 1). However, the proportion of sequences present in 10 copies or more was not significantly affected (Fig. 2). When SD was set at 1 so that one unit could diverge each time that a duplication or deletion event occurred, 17% of the classes were present in three copies or less and 63% were present in 10 copies or more. At higher SD rates the percent repetitive sequence and mean copy number decreased (Fig. 2). The characteristics of the genomes generated at these rates of sequence divergence resemble those of many organisms with respect to repetitive DNA (3).

Fig. 2. Proportions of repetitive DNA. The percent of total sequences present in 10 copies or more (solid symbols) was calculated in genomes generated at different SD settings. The mean copy number (open symbols) was calculated in the same genomes after 10,000 events (circles) and again after 20,000 events (squares). There was less than 10% variation in independent simulations.
Genome Size

Certain eukaryotic organisms, such as the mold Aspergillus, have very little repetitive DNA (7). The genomes in these organisms are all quite small. To simulate such genomes the frequency of duplications was reduced and the frequency of deletions increased for 10,000 events. The smallest genomes were found to have only 1% of their sequences present in more than 10 copies (Fig. 3). The percent repetitive sequences increased dramatically as the genome was allowed to expand. When duplications occurred in 70% of the events and sequence divergences occurred at the same frequency as duplication/deletion events (SD = 1), the genome grew to 44,778 units and 68% of the sequences were present in repetitive families that had a mean of 74 ± 6 members. A quarter of the sequences were present in families with more than 80 members. It made little difference if the rate of sequence divergence was doubled; however, in the absence of any sequence divergence even small genomes carried a high proportion of repetitive sequences and this proportion increased in large genomes.

Fig. 3. Size dependence. Genomes of different sizes were generated by varying the proportion of duplications per event from 0.2 to 0.7 for 10,000 events. The percent of the total number of classes that were present in 10 or more copies was determined in each genome. SD was set at 0 (closed boxes), 1 (open circles) or 2 (closed circles). SAC was set at 0.1/G.

Families of Repetitive Sequences

The process of genetic duplication followed by sequence divergence generates interspersed repetitive families in the genome. The average size of the families varies from 5 to several hundred depending on the rate of divergence, the size of duplications, and the length of the genome. The GENESIS program produces genomes representative of those observed in a wide variety of organisms. The program relies on random duplications and deletions in conjunction with divergence rates that seem plausible. Up to 70% of the genome of many organisms consists of interspersed repetitive DNA present in families with 10 to 100 members than can be accounted for by the processes simulated by GENESIS. They also carry repetitive DNA present in much higher copy number that is often localized such that it can be isolated as satellite bands upon density separation. These highly repeated sequences appear to be
Neutral Mutations and Repetitive DNA

generated by unequal crossing over within localized regions (10). Our analysis of
genome evolution does not include such sequences or those generated by self-
replicating sequences and is pertinent only to the middle repetitive class of DNA.

Sequence divergence in our model is dependent on the presence of dispensable
portions of the genome. These are generated as a consequence of the restriction of
acceptable deletions to portions that do not contain the sole remaining copy of a gene.
Duplications, on the other hand, are acceptable even if they include only a non-
functional portion of a gene. In our previous study, we showed that only 4% of a
complex genome of 13,710 units generated by these rules consisted of vital sequences
(9). Direct analysis by insertional mutagenesis of *Saccharomyces cerevisiae* has since
shown that less than 15% of the 13,000 kilobase yeast genome is essential for growth
and development under laboratory conditions (14).

The SD values found to give the best fit to observed proportions of repetitive
DNA are dependent on the maximum size of the duplication/delition events and the
length of the genome generated. Since it is difficult to arrive at independent estimates
for these parameters in different organisms, undue weight should not be placed on the
specific SD values. However, it is clear that sequence divergence is innate in replication
of genomes over evolutionary time scales and that it is essential in our simulations to
generate representative genomes.

Organisms with small genomes may have been under selective pressure to trim
their genomes. Alternatively, in these organisms, the mechanism of replication of the
genome is more subject to deletion events than to duplication events. Our simulations of
such genomes indicate that the lower proportion of repetitive DNA in small genomes is
dependent on sequence divergence occurring at a frequency close to that of
duplication/deletion events.

The copy number of individual genes fluctuates as genomes evolve in our
simulations. Most genes present in multiple copies at the end of 10,000 events were
generated from a single copy within the last few thousand events. Therefore, they are
more closely related to each other than they are to the same gene in an independent
genome. Such a “founder-effect” may account for the observation that members of
multigene families are often more similar to each other than they are to homologous
genes in separate species. This rectification mechanism would affect repetitive vestigial
sequences as well as functional genes. Although unequal crossing-over during meiotic
recombination is one of mechanisms that generates both duplications and deletions in
diploid organisms, other mechanisms appear to function in haploid organisms since
they also carry homogenized multigene families (9). Thus, rectification need not
depend on unequal crossing-over but will result from any process that randomly
generates duplications and deletions.

For the last 15 years, the function of middle repetitive DNA has been the subject of
considerable speculation (1, 3). It has been proposed that some of the repeated
sequences may be cis-acting control regions recognized by regulatory proteins to effect
coordinate expression of physiologically related genes. However, most cis-acting
sequences have been found to be less than 20 bases long and so all together would make
up only a small proportion of the middle repetitive DNA even if most genes shared
common control regions. Moreover, such control regions are too short to reanneal
under standard conditions. Our simulations suggest that most middle repetitive
sequences are unavoidable products of duplication and divergence of inactive portions of genes that serve no purpose but are too insignificant to be removed from genomes.

REFERENCES