Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 2143–2147, April 1986
Evolution

Multigene families and vestigial sequences

WILLIAM F. LOOMIS AND MICHAEL E. GILPIN

Department of Biology, University of California at San Diego, La Jolla, CA 92093

Communicated by Russell F. Doolittle, November 12, 1985

ABSTRACT **Random duplication and deletion events generate complex genomes carrying a large amount of dispensable sequences. We have simulated such events in a computer model. We followed the evolution of a genome carrying at least one copy of each type of gene. Partial duplications and deletions of genes generated nonfunctional vestigial sequences that were dispensable. The size of the genome stabilized only when the amount of dispensable sequences had increased to the point that most deletions did not affect vital genes. Within such genomes, the number of copies of specific genes fluctuated, thereby generating small multigene families. The parameters of the model were tested over 100,000 events in both simple and complex genomes. The results indicate that when the size of the genome is not critical to survival, as appears to be the case within limits in most eukaryotic organisms, the genome carries vestigial sequences that are no longer functional and that many genes are present in multigene families by chance.**

Evolution of complex genomes from simpler ones appears to have involved duplication of individual genes and surrounding sequences followed by changes in the extra copies. In some instances the common ancestry of genes or portions of genes can be recognized by comparison of amino acid and nucleotide sequences (1). Many related DNA sequences, such as those coding for α- and β-globins, produce proteins with related functions, while other related sequences have sustained deleterious mutations or deletions such that they are no longer functional and are referred to as pseudogenes (2). The number of members in a specific multigene family, such as the actin family (3), varies widely between genomes of different organisms. There is no obvious explanation for this variability.

Another feature of the genomes of eukaryotic organisms is a large excess in potential coding capacity carried in unique DNA sequences that far exceeds the estimated amount under selective surveillance (4). The amount of unique sequences in the total DNA of different genomes varies widely between even closely related species, although the number of selectively advantageous genes is likely to lie within a fairly narrow range. Some of the excess single-copy sequences have been found to be generated by retrovirus and retrotransposon integration into the genome (5). These sequences have been considered "selfish DNA" (6, 7). However, they appear to make up only a small fraction of the unaccounted sequences (5). The presence of introns within genes requires transcribed regions to be larger than had been estimated previously from mRNA size. Nevertheless, only a small percentage of the genome is transcribed at any stage in the life cycle of eukaryotes (8). The great majority of single-copy DNA sequences does not appear to serve any selectively advantageous function and may be dispensable.

Duplications within a genome are seldom detrimental, but deletions that remove a selectively advantageous sequence will ultimately result in the loss of that genome from the population. The end points for duplications and deletions appear to be random and fail to distinguish between transcribed and dispensable sequences. Inactivation of an extra copy of a gene by partial duplication or deletion produces vestigial DNA sequences.

The size of a genome is increased by duplications of sequences and reduced by deletions. Equilibrium will be reached when the amount of dispensable DNA results in the rate of nondeleterious deletions being equal to the rate of duplications (9). Target theory requires that a large proportion of a genome near equilibrium size be dispensable unless there is selective advantage for a small genome. In viruses the genome must be packaged in small particles, and in bacteria the maximum growth rate is limited by the time required for replication of the genome. In these organisms there appears to be a selective advantage for small genomes with fewer dispensable sequences. On the other hand, eukaryotes have no difficulty packaging large amounts of DNA in nuclei, and the cell cycle period is greater than the period when DNA is replicated (S phase). Thus, there is no apparent restraint on the present-day size of the genome in eukaryotes. Although it is inefficient to replicate and carry dispensable sequences, the energy expended in DNA replication can be calculated to be a negligible load on a cell's economy (9).

We have simulated the processes of duplication and deletion on simple genomes in a computer model. Using only a few plausible rules, we find that a genome will approach equilibrium with a fairly large amount of dispensable sequences and will carry a fluctuating number of copies of a given gene. There is often only a single copy of an essential gene, but at times duplications will generate small multigene families even though a single copy would be sufficient. Thus, the occurrence of multigene families as well as excess DNA is a consequence of random duplications and deletions.

SIMULATION

Genomic evolution was simulated on a VAX computer. Each event consisted of a randomly chosen duplication or deletion of 1–20 units. The position of the event was also randomly generated within the genome. When a duplication occurred, the new sequence was positioned next to the existing sequence. A vital gene was represented as 4 units consisting of 1 start unit (promoter), 2 coding units, and 1 stop unit (terminator). The initial genome contained only a single gene represented by "#AA." (see Fig. 1).

Duplications that included all four units of a gene resulted in two equivalent copies of the gene. Duplications that only included a portion of a gene resulting in a nonfunctional sequence were represented by dashes. Deletions that removed a portion of a duplicated gene, thereby rendering it nonfunctional, converted the remaining units into vestigial sequences, also represented by dashes. Deletions in only a single copy were considered lethal and terminated the evolution of that genome. The pedigree was then continued

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: SAC, selectively advantageous change; kb, kilobase(s).

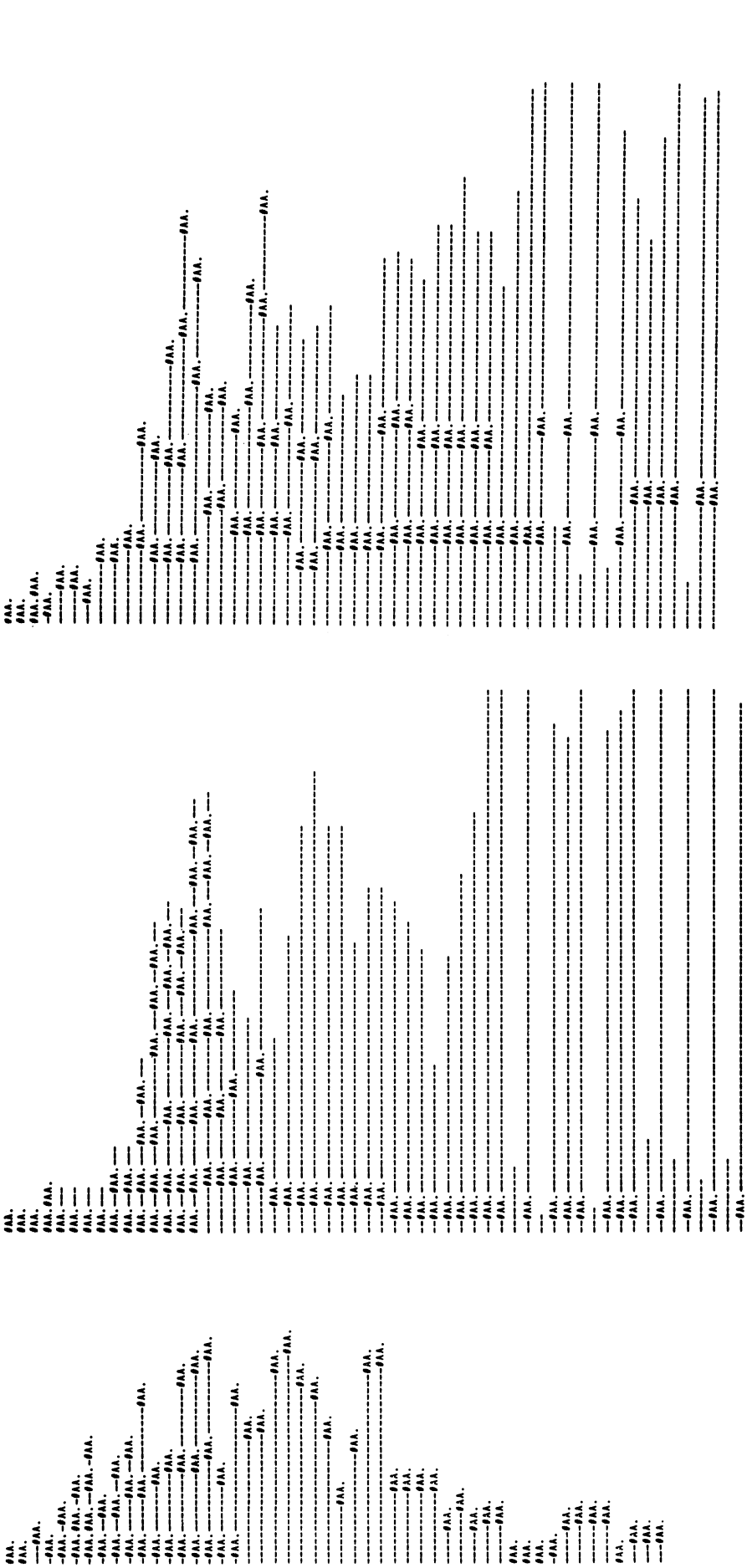

FIG. 1. Early events in genomic evolution. Three independent simulations were carried out for 50 events. The essential gene (#AA.) was duplicated to give a small multigene family. Duplication or deletions of less than the complete 4 units of a gene resulted in vestigial sequences (----). The exact event can be deduced by analysis of the genomes before and after the event. The probability of a duplication event or a deletion event was set equal to each other. The number of units affected was randomly chosen between 1 and 20. The end points were also randomly selected.

from the preceding viable genome. The consequences of up to 100,000 random events were followed.

To simulate evolution of genetic complexity, genes present in multiple copies were given a probability of changing to new and different genes. These were represented as two types of change: (*i*) capital letter changes (e.g., #AA. to #AB.) and (*ii*) lower-case letter changes (e.g., #AA. to #Aa. or #Db. to #db.). Genes with lower-case letters could then evolve along the alphabet in lower case (e.g., #Aa. to #Ab.). This representation can distinguish several thousand genes. Only a gene present in two or more copies at a given state could change to a new gene. In this way an essential gene was never lost from the genome. The first member of a multigene family was chosen for the genetic change. Only the genome carrying the greatest number of different selectively advantageous genes was followed.

DISPENSABLE SEQUENCES

Starting with a genome containing only the four units representing a gene (#AA.), random duplications and deletions generated genomes containing up to 100 units within 50 events (Fig. 1). The size of the genome fluctuated widely during the early events, as might be expected when the size of duplications or deletions could be a significant proportion of the genomic length. At later stages in the simulation, when the genome had increased to several thousand units, the relative fluctuation was much smaller (Fig. 2). Likewise, the number of copies of the gene initially fluctuated rapidly but then settled to a lower rate of fluctuation (Figs. 1 and 2). The average number of functional copies of the gene over 30,000 events was 1.67 as a consequence of often being present as a single copy. After 1000 events the genome had expanded to over 100 units. The size increased to about 5000 units after 20,000 events and then changed very little because most deletions were not deleterious in that they occurred within the dispensable sequences derived from incomplete portions of the gene. On average, the nonlethal deletion rate almost equaled the duplication rate, since only 0.4% of the genome was essential when there were 1000 units. When the rate of duplication was set greater than the rate of deletion, the genome expanded rapidly without stabilizing. When the rate of duplication was set less than the rate of deletion, the genome remained small. However, in neither case was the mean number of functional copies of the gene significantly affected, since in large genomes the probability of a duplication or deletion occurring in the region of the gene was lower than in small genomes. When the maximum size of a duplication or deletion event was increased to more than 20 units, there was an exaggeration of local clustering of genes or dispensable sequences at the early stages of the simulation, but there was little effect after 10,000 events.

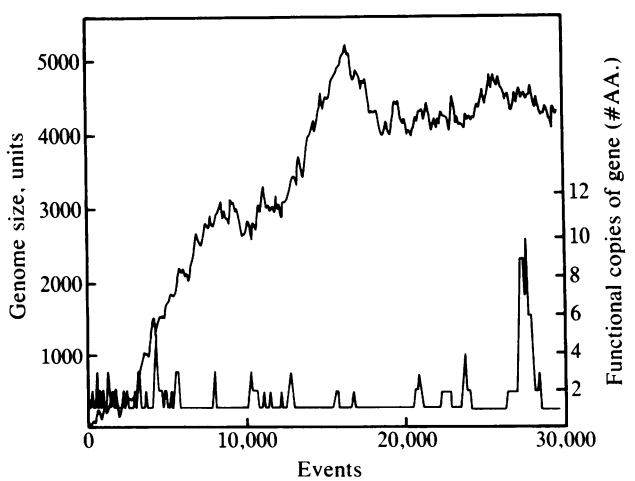

FIG. 2. Fluctuations in family genome size (upper trace). A simulation similar to the ones carried out in Fig. 1 was followed for 30,000 events. The number of units in the total genome increased to several thousand. In longer simulations (100,000 events), genomes containing up to 8000 units were generated. The number of functional copies of the gene (lower trace) varied from 1 to 10 with a mean of 1.67.

COMPLEX GENOMES

Actual genomes carry many different genes. To generate more realistic genomes, we allowed the initial gene (#AA.) to change in either of two ways to generate new genes. At various frequencies, a change could occur represented by an upper-case letter in alphabetical order. Thus, #AA. could change to #AB.; a 10-fold less probable change was represented by a change from upper-case to lower-case letter such that #AB. could change to #Ba. on a random basis. In all cases a change could affect only a gene present in more than one copy so as to leave at least one copy of the original gene. The new gene was considered to be selectively advantageous; thus, only the genome with the greatest number of different genes was followed as a simulation of adaptation. Since an organism with few genes would be expected to benefit more often from random changes in genes than would an organism already carrying many genes, the selectively advantageous change (SAC) rate per event was made inversely proportional to the number of different genes (G) in the genome. This results in a linear rate of generation of new genes as events proceed in these primitive genomes.

When the SAC rate per event was set at $0.1/G$, a genome of 6356 units containing 83 different genes was generated in 10,000 events (Fig. 3). Each gene was present in 1.58 copies on the average, although there were 54 single-copy genes (Table 1). There were 7 copies of one gene (#db.). The genes were spread out over most of the genome, although a long uninterrupted segment of dispensable sequences occurred in the first half.

After another 10,000 events, this genome had increased to 13,710 units carrying 124 genes. The average size of multigene families was 1.67. After 30,000 events the genome was 18,120 units long and had 165 different genes. A new gene was generated about every 180 events, since a given gene family was randomly selected every 10 events, but less than 10% of the genes were present in multiple copies that could undergo a SAC. The long segment of dispensable sequences had been retained. The size of the genome increased linearly in parallel with the number of different genes. The proportion of these genomes carrying essential genetic information was 3.6%. This is higher than the genomes carrying only a single type of gene (Fig. 2), partly because the greater number of genes reduced the number of acceptable deletions and partly because of the high SAC rate we set to generate complexity. To simulate more evolved genomes, we stopped any further SAC when the genomes contained various numbers of different genes and allowed duplication and deletion events to proceed (Fig. 4). In all cases the proportion of the genome carrying essential genetic information was reduced to less than 2% when 30,000 events occurred after setting the SAC rate to zero.

DISCUSSION

When the potential coding capacity of the genomes of a large number of species are compared, they are found to vary widely (8). It is not clear why such closely related organisms as anuran and urodele amphibians should differ in the amount of single-copy DNA complexity by a factor of 5 and that both should have more than 10 times the potential coding capacity

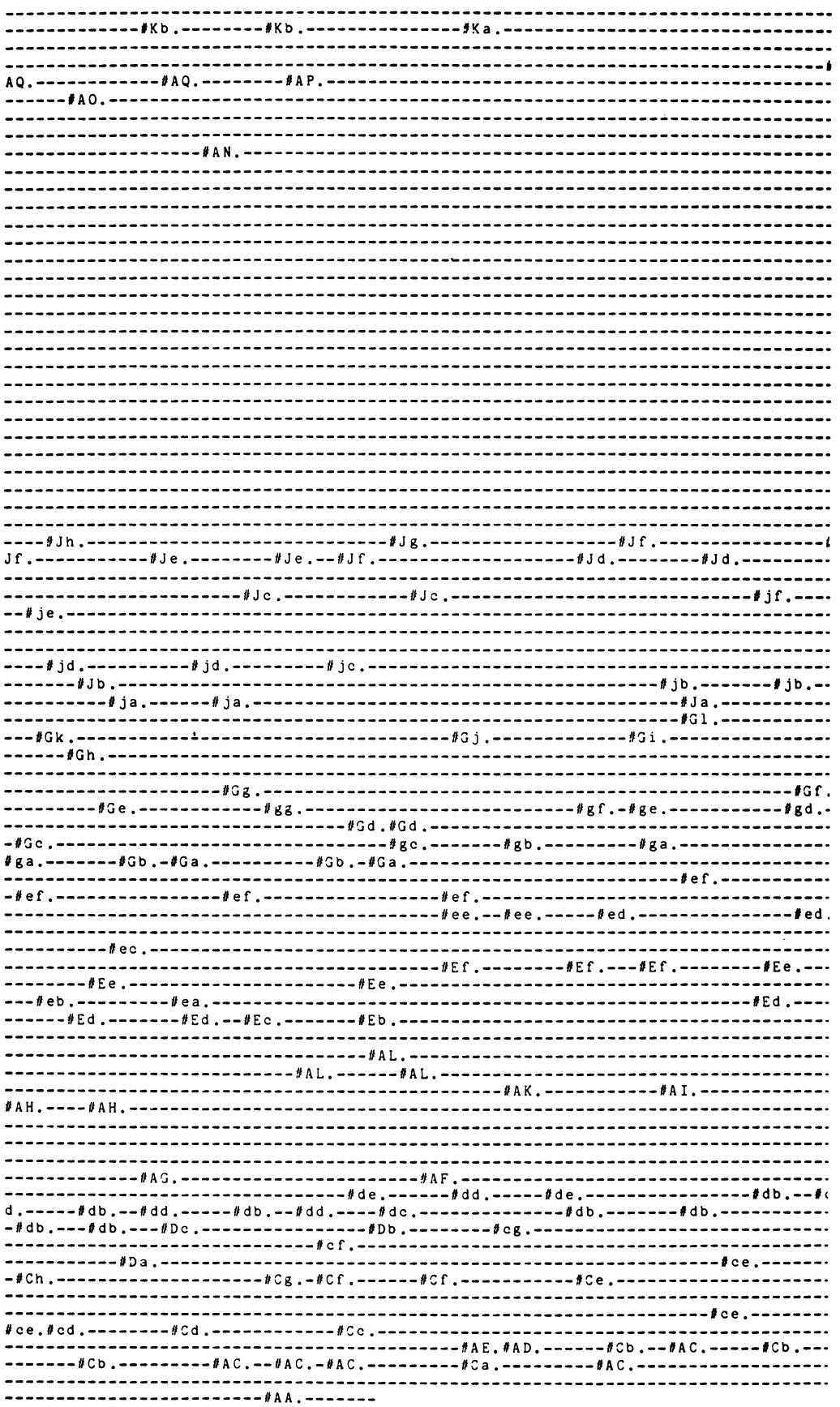

FIG. 3. Complex genome. After 10,000 events, 83 different genes were generated by setting the frequency at which multicopy genes could undergo a SAC at $0.1/G$ (number of different genes). This SAC rate per event occurred for both upper-case and lower-case letter changes. The rate of upper- to lower-case letter changes was set at 1/10th this rate. Duplication or deletions occurred at each event. Dashes are vestigial sequences.

Table 1. Gene frequencies

Gene	Number	Gene	Number	Gene	Number	Gene	Number
AA	1	Cc	1	AN	1	Gh	1
AB	1	Gb	2	jd	2	AP	1
AC	5	Jb	1	cf	1	ge	1
AD	1	Cd	1	db	7	gf	1
AE	1	Ed	3	Cf	2	ga	2
AF	1	AL	3	Ge	1	dd	4
AG	1	ec	1	Ee	3	cg	1
AH	2	ed	2	Jf	3	Ef	3
Ea	1	Jc	2	Gf	1	Jh	1
AI	1	ee	2	je	1	Gi	1
AJ	1	Ce	1	Gg	1	gb	1
ea	1	Jd	2	ja	2	de	2
Eb	1	Gc	1	Cg	1	Gj	1
Ca	1	AM	1	gd	1	Gk	1
Ga	2	Je	2	dc	1	gc	1
AK	1	Gd	2	cd	2	AQ	2
Cb	3	ce	3	Dc	1	Gl	1
Ka	1	Da	1	AO	1	Ch	1
Ja	1	jb	2	Jg	1	gg	1
eb	1	jc	1	jf	1	Kb	2
Ec	1	Db	1	ef	4		

of humans. Mammals have sufficient DNA information to code for 5×10^5 transcripts of 4 kilobases (kb) each. However, direct measurement of the total number of different mRNAs gives less than 5×10^4 transcripts (8). These studies have relied on global rehybridization analyses of DNA and RNA.

More recently, large portions of the DNA surrounding a variety of cloned genes from diverse organisms have been characterized, and many have been sequenced. Most of the time, no long open-reading frames that could potentially carry genetic information were found within several kilobases of a specific gene. Most of the flanking sequences are functionally dispensable when tested by transformation of the DNA back into cells. Molecular maps covering 100 kb of DNA are characterized by islands of transcribed sequences in a sea of silent DNA.

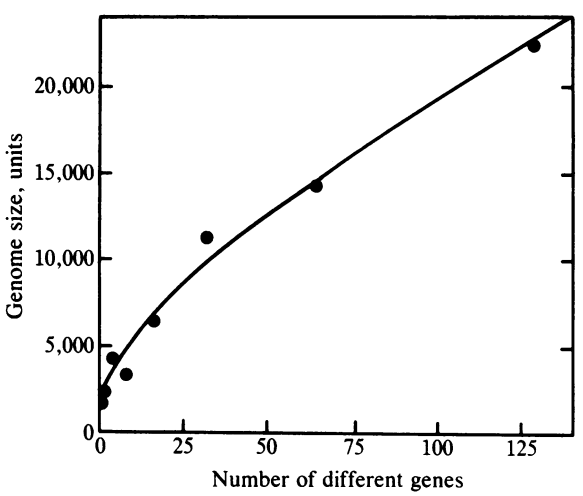

FIG. 4. Increase in genome size. Genomes with 1–128 different genes were produced by the random duplication/deletion program with SAC=0.1/G. Further generation of new types of genes was then stopped (SAC = 0), and the program was run for an additional 30,000 events. Three independent evolutionary simulations were run, and the final genome sizes were averaged. The variability in genome size was greater in simple genomes than in the more complex ones. The size of genomes carrying 16 or more different genes varied by less than 10% in independent simulations.

Our analysis of random duplications and deletions under selective surveillance to eliminate deleterious deletions shows that these processes by themselves require that most of the DNA be dispensable. Genes duplicated as the genome expanded during evolution. Thus, most DNA sequences were once members of multigene families. Duplication of an incomplete portion of a gene resulted in vestigial sequences because they no longer carried the full functional information. Although we have not considered point mutations that inactivate genes in our simulations, they are clearly a major cause of loss of function of duplicated genes. When a gene has sustained such a mutation recently in evolution, the sequence of this vestigial DNA is sufficiently similar to that of the functioning gene that it is referred to as a pseudogene (2). Clustering of similar pseudogenes indicates that these vestigial sequences also are subject to duplication and deletion events. A pseudogene is free to mutate randomly and will diverge rapidly until its sequence no longer resembles that of the ancestral gene. The vestigial DNA then appears to be a random sequence of bases.

It is a curious fact that conserved genes such as actin vary in the number of copies from 17 in the slime mold *Dictyostelium discoideum* to 1 in the yeast *Saccharomyces cerevisiae* (3). The genome of the fly *Drosophila melanogaster* and that of the sea urchin *Strongylocentrotus purpuratus* carry 6 actin genes. There seems no reason why the size of this multigene family should vary so. Our analysis of random duplications and deletions suggests that multigene families fluctuate over this range when chance events encompass a member or members of the family.

In our simulation, a gene was represented by 4 units. If each unit is 1 kb of DNA, then genomes of 18,000 units, such as the one with 165 genes generated in 30,000 events (Fig. 3), would have 18,000 kb of DNA. Those of 10^5 to 10^6 units would be of the size found in eukaryotes and would carry up to 10,000 different genes.

Replication of genomes is highly accurate in present-day organisms but may have been subject to more error early in evolution. If a duplication or deletion event occurred once every 10^3 generations, our simulations that extended to 10^5 events would span 10^8 generations. These would not require a high proportion of the span of evolutionary time. With the advent of chromosomal recombination, the possibility of unequal crossing-over would increase the rate of duplications and deletions and would serve to disperse members of multigene families.

1. Doolittle, R. F. (1985) *Trends Biochem. Sci.* **10**, 233–237.
2. Proudfoot, N. (1980) *Nature (London)* **286**, 840–841.
3. Firtel, R. A. (1981) *Cell* **24**, 6–7.
4. Ohta, R. & Kimura, M. (1971) *Nature (London)* **233**, 118–119.
5. Baltimore, D. (1985) *Cell* **40**, 481–482.
6. Doolittle, W. F. & Sapienza, C. (1980) *Nature (London)* **284**, 601–603.
7. Orgel, L. & Crick, F. H. C. (1980) *Nature (London)* **284**, 604–607.
8. Davidson, E. (1976) *Gene Activity in Early Development* (Academic, New York).
9. Loomis, W. F. (1973) *Dev. Biol.* **30**, f.3–f.4.