Selfish DNA: the ultimate parasite

L. E. Orgel & F. H. C. Crick
The Salk Institute, 10010 N. Torrey Pines Road, La Jolla, California 92037

The DNA of higher organisms usually falls into two classes, one specific and the other comparatively nonspecific. It seems plausible that most of the latter originated by the spreading of sequences which had little or no effect on the phenotype. We examine this idea from the point of view of the natural selection of preferred replicators within the genome.

THE object of this short review is to make widely known the idea of selfish DNA. A piece of selfish DNA, in its purest form, has two distinct properties:

(1) It arises when a DNA sequence spreads by forming additional copies of itself within the genome.

(2) It makes no specific contribution to the phenotype.

This idea is not new. We have not attempted to trace it back to its roots. It is sketched briefly but clearly by Dawkins[1] in his book *The Selfish Gene* (page 47). The extended discussion (pages 39–45) after P. M. B. Walker's article[2] in the CIBA volume based on a Symposium on Human Genetics held in June 1978 shows that it was at that time already familiar to Bodmer, Fincham and one of us. That discussion referred specifically to repetitive DNA because that was the topic of Walker's article, but we shall use the term selfish DNA in a wider sense, so that it can refer not only to obviously repetitive DNA but also to certain other DNA sequences which appear to have little or no function, such as much of the DNA in the introns of genes and parts of the DNA sequences between genes. The catch-phrase 'selfish DNA' has already been mentioned briefly on two occasions[3,4]. Doolittle and Sapienza[5] (see the previous article) have independently arrived at similar ideas.

The amount of DNA

The large amounts of DNA in the cells of most higher organisms and, in particular, the exceptionally large amounts in certain animal and plant species—the so-called *C* value paradox—has been an unsolved puzzle for a considerable period (see reviews in refs 6–8). As is well known, this DNA consists in part of 'simple' sequences, an extreme example of which is the very large amounts of fairly pure poly d(AT) in certain crabs. Simple sequences, which are situated in chromosomes largely but not entirely in the heterochromatin, are usually not transcribed. Another class of repetitive sequences, the so-called 'intermediate repetitive', have much longer and less regular repeats. Such sequences are interspersed with 'unique' DNA at many places in the chromosome, the precise pattern of interspersion being to some extent different in different species. Leaving aside genes which code for structural RNA of one sort or another (such as transfer RNA and ribosomal RNA), which would be expected to occur in multiple copies (since, unlike protein, their final products are the result of only one stage of magnification, not two), the majority of genes coding for proteins appear to exist in 'single' copies, meaning here one or a few. A typical example would be the genes for α-globin, which occur in one to three copies and the human β-like globins, of which there are four main types, all related to each other but used for slightly different purposes. Notable exceptions are the proteins of the immune system, and probably those of the histocompatibility and related systems. Another exception is the genes for the five major types of histone which also occur in multiple copies. Even allowing for all such special case, the estimated number of genes in the human genome appears too few to account for the 3×10^9 base pairs found per haploid set of DNA, although it must be admitted that all such arguments are very far from conclusive.

Several authors[8–13] have suggested that the DNA of higher organisms consists of a minority of sequences with highly specific functions plus a majority with little or no specificity. Even some of the so-called single-copy DNA may have no specific function. A striking example comes from the study of two rather similar species of *Xenopus*. These can form viable hybrids, although these hybrids are usually sterile. However, detailed molecular hybridization studies show that there has been a large amount of DNA sequence divergence since the evolutionary separation of their forebears. These authors[13] conclude 'only one interpretation seems reasonable, and that is that the specific sequence of much of the single-copy DNA is not functionally required during the life of the animal. This is not to say that this DNA is functionless, only that its specific sequence is not important'.

There is also evidence to suggest that the majority of DNA sequences in most higher organisms do not code for protein since they do not occur at all in messenger RNA (for reviews see refs 14, 15). Nor is it very plausible that all this extra DNA is needed for gene control, although some portion of it certainly must be.

We also have to account for the vast amount of DNA found in certain species, such as lilies and salamanders, which may amount to as much as 20 times that found in the human genome. It seems totally implausible that the number of radically different genes needed in a salamander is 20 times that in a man. Nor is there evidence to support the idea that salamander genes are mostly present in about 20 fairly similar copies. The conviction has been growing that much of this extra DNA is 'junk', in other words, that it has little specificity and conveys little or no selective advantage to the organism.

Another place where there appears to be more nucleic acid than one might expect is in the primary transcripts of the DNA of higher organisms which are found in the so-called heteronuclear RNA. It has been known for some time that this RNA is typically longer than the messenger RNA molecules found in the corresponding cytoplasm. Heteronuclear RNA contains these messenger RNA sequences but has many other sequences which are never found in the cytoplasm. The phenomenon has been somewhat clarified by the recent discovery of introns in many genes (for a general introduction see ref. 4). Although the evidence is still very preliminary, it certainly suggests that much of the base sequence in the interior of some introns may be junk, in that these sequences drift rapidly in evolution, both in detail and in size. Moreover, the number of introns may differ even in closely related genes, as in the two genes for rat preproinsulin[16]. Whether there is junk between genes is unclear but it is noteworthy that the four genes for the human β-like globins, which occur fairly near together in a single stretch of DNA, occupy a region no less than 40 kilobases long[17]. This greatly exceeds the total length of the four primary transcripts (that is the four mRNA precursors), an amount estimated to be considerably less than 10 kilobases. There is little evidence to indicate that there are other coding sequences between these genes (although the question is still quite open) and a tenable hypothesis is that much of this interspersed DNA has little specific function.

In summary, then, there is a large amount of evidence which suggests, but does not prove, that much DNA in higher organisms is little better than junk. We shall assume, for the rest of this article, that this hypothesis is true. We therefore need to explain how such DNA arose in the first place and why it is not speedily eliminated, since, by definition, it contributes little or nothing to the fitness of the organism.

What is selfish DNA?

The theory of natural selection, in its more general formulation, deals with the competition between replicating entities. It shows that, in such a competition, the more efficient replicators increase in number at the expense of their less efficient competitors. After a sufficient time, only the most efficient replicators survive. The idea of selfish DNA is firmly based on this general theory of natural selection, but it deals with selection in an unfamiliar context.

The familiar neo-darwinian theory of natural selection is concerned with the competition between organisms in a population. At the level of molecular genetics it provides an explanation of the spread of 'useful' genes or DNA sequences within a population. Organisms that carry a gene that contributes positively to fitness tend to increase their representation at the expense of organisms lacking that gene. In time, only those organisms that carry the useful gene survive. Natural selection also predicts the spread of a gene or other DNA sequence within a single genome, provided certain conditions are satisfied. If an organism carrying several copies of the sequence is fitter than an organism carrying a single copy, and if mechanisms exist for the multiplication of the relevant sequence, then natural selection must lead to the emergence of a population in which the sequence is represented several times in every genome.

The idea of selfish DNA is different. It is again concerned with the spread of a given DNA within the genome. However, in the case of selfish DNA, the sequence which spreads makes no contribution to the phenotype of the organism, except insofar as it is a slight burden to the cell that contains it. Selfish DNA sequences may be transcribed in some cases and not in others. The spread of selfish DNA sequences within the genome can be compared to the spread of a not-too-harmful parasite within its host.

Mechanisms for DNA spreading

The inheritance of a repeated DNA sequence in a population of eukaryotes clearly requires that the multiplication which produced it occurred in the germ line. Furthermore, any mechanism that can lead to the multiplication of useful DNA will probably lead to the multiplication of selfish DNA (and vice versa). Of course, natural selection subsequently discriminates between multiple sequences of different kinds, but it does not necessarily prevent the multiplication of neutral or harmful sequences.

Multiplication in the germ-line sequence can occur in nondividing cells or during meiosis and mitosis (within lineages that lead to the germ line). In the former case, the mechanisms available resemble those that are well documented for prokaryotes, that is, multiplication may occur in eukaryotes through the integration of viruses or of elements analogous to transposons and insertion sequences. Doolittle and Sapienza[5] have discussed these mechanisms in some detail, particularly for prokaryotes. They are likely to lead to the spreading of DNA sequences to widely separated positions on the chromosomes.

During mitosis and meiosis, multiplication (or deletion) is likely to occur by unequal crossing over. This mechanism will often lead to the formation of tandem repeats. It is well documented for the tRNA 'genes' of *Drosophila* and for various other tandemly repeated sequences in higher organisms.

The amount and location of selfish DNA

Natural selection 'within' the genome will favour the indefinite spreading of selfish preferred replicators. Natural selection between genotypes provides a balancing force that attempts to maintain the total amount of selfish DNA at an equilibrium (steady state) level—organisms whose genomes contain an excessive proportion of selfish DNA would be at a metabolic disadvantage relative to organisms with less selfish DNA, and so would be eliminated by the normal mechanism of natural selection. Excessive spreading of functionless replicators may be considered as a 'cancer' of the genome—the uncontrolled expansion of one segment of the genome would ultimately lead to the extinction of the genotype that permits such expansion. Of course, we do not know whether extinction of genotypes in nature even occurs for this reason.

It is hard to get beyond generalities of this kind. To do so we would, at least, need to know how much selective disadvantage results from the presence of a given amount of useless DNA. Even this minimal information is not easily acquired, so we cannot produce other than qualitative arguments.

It seems certain that the metabolic energy cost of replicating a superfluous short DNA sequence in a genome containing 10^9 base pairs would be very small. If, for example, the selective advantage were equal to the proportion of the genome made up by the extra DNA, a sequence of 1,000 base pairs would produce a selective disadvantage of only 10^{-6}. If the selective disadvantage were proportional to the extra energy cost divided by the total metabolic energy expended per cell per generation, the disadvantage would be much smaller. The selective disadvantage might be greater in more stringent conditions, but it is still hard to believe that a relatively small proportion of selfish DNA could be selected against strongly.

On the other hand, when the total amount of selfish DNA becomes comparable to or greater than that of useful DNA, it seems likely that the selective disadvantage would be significant. We may expect, therefore, that the mechanisms for the formation and deletion of nonspecific DNA will adjust, in each organism, so that the load of DNA is sufficiently small that it can be accommodated without producing a large selective disadvantage. The proportion of nonspecific DNA in any particular organism will thus depend on the lifestyle of the organism, and particularly on its sensitivity to metabolic stress during the most vulnerable part of the life cycle.

We can make one prediction on the basis of energy costs. Selfish DNA will accumulate to a greater extent in non-transcribed regions of the genome than in those that are transcribed. Of course, selfish DNA will in most cases be excluded from translated sequences, because the insertion of amino acids within a protein will almost always have serious consequences, even in diploid organisms (but see the suggestion by F.H.C.C.[18]).

At first sight it might seem anomalous that natural selection does not eliminate all selfish DNA. Since the suggestion that much eukaryotic DNA is useless distinguishes the selfish DNA hypothesis from many closely related proposals, it may be useful to take up this point in some detail.

First, the elimination of disadvantaged organisms from a population, by their more favoured competitors, takes a number of generations several times larger than the reciprocal of the selective disadvantage. If the selective disadvantage associated with a stretch of useless DNA in higher organisms is only 10^{-6}, it would take 10^6–10^8 years to eliminate it by competition. For typical higher organisms this is a very long time, so the elimination of a particular stretch of selfish DNA may be a very slow process even on a geological time scale. Second, the mechanisms for the deletion of short sequences of DNA may be inefficient, since there is no strong selective pressure for the development of 'corrective' measures when the 'fault' carries a relatively small selective penalty. Taken together, these arguments suggest that the elimination of a particular piece of junk from the genome may be a very slow process.

This in turn suggests that the amount of useless DNA in the genome is a consequence of a dynamic balance. The organism 'attempts' to limit the spread of selfish DNA by controlling the mechanism for gene duplication, but is constrained by imperfections in genetic processes and/or by the need to permit some duplication of advantageous genes. Selfish DNA sequences 'attempt' to subvert these mechanisms and may be able to do so comparatively rapidly because mutation will affect them directly. On the other hand, the defence mechanisms of the host are likely to depend on the action of protein and therefore may evolve more slowly. Once established within the genome, useless sequences probably have a long 'life expectancy'.

For any particular type of selfish DNA, there is no reason that a steady state should necessarily be reached in evolution. The situation would be continually changing. A particular type of DNA might first spread rather successfully over the chromosomes. The host might then evolve a mechanism which reduced or eliminated further spreading. It might also evolve a method for preferentially deleting it. At the same time, random mutations in the selfish DNA might make it more like ordinary DNA and so, perhaps, less easy to remove. Eventually, these sequences, possibly by now rather remote from those originally introduced, may cease to spread and be slowly eliminated. Meanwhile, other types of selfish DNA may originate, expand and evolve in a similar way.

In short, we may expect a kind of molecular struggle for existence within the DNA of the chromosomes, using the process of natural selection. There is no reason to believe that this is likely to be much simpler or more easy to predict than evolution at any other level. At bottom, the existence of selfish DNA is possible because DNA is a molecule which is replicated very easily and because selfish DNA occurs in an environment in which DNA replication is a necessity. It thus has the opportunity of subverting these essential mechanisms to its own purpose.

The inheritance of selfish DNA

Although the inheritance of selfish DNA will occur mainly within a mendelian framework, it is likely to be different in detail and more complex than simple mendelian inheritance. This is due both to the multiplication mechanisms, which in one way or another will produce repeated copies (see the discussion by Doolittle and Sapienza[5]), and to the fact that these copies are likely to be distributed round the chromosomes rather than being located in a single place in the genome as most normal genes are. For both these reasons, a particular type of selfish DNA is likely to spread more rapidly through a population than would a normal gene with a low selective advantage. It will be even more rapid if selfish DNA can spread horizontally between different individuals in a population, due to viruses or other infectious agents, although it should be remembered that such 'infection' must affect the germ line and not merely the soma. If this initial spread takes place when the additional DNA produced is relatively small in amount, it is unlikely to be seriously hindered by the organism selecting against it. The study of these processes will clearly require a new type of population genetics.

Can selfish DNA acquire a specific function?

It would be surprising if the host organism did not occasionally find some use for particular selfish DNA sequences, especially if there were many different sequences widely distributed over the chromosomes. One obvious use, as repeatedly stressed by Britten and Davidson[19,20], would be for control purposes at one level or another. This seems more than plausible.

It has often been argued (see, for example, ref. 21) that for the evolution of complex higher organisms, what is required is not so much the evolution of new proteins as the evolution of new control mechanisms and especially mechanisms which control together sets of genes which previously had been regulated separately. To be useful, a new control sequence on the DNA is likely to be needed in a number of distinct places in the genome. It has rarely been considered how this could be brought about expeditiously by the rather random methods available to natural selection.

A mechanism which scattered, more or less at random, many kinds of repeated sequences in many places in the genome would appear to be rather good for this purpose. Most sets of such sequences would be unlikely to find themselves in the right combination of places to be useful but, by chance, the members of one particular set might be located so that they could be used to turn on (or turn off) together a set of genes which had never been controlled before in a coordinated way. A next way of doing this would be to use as control sequences not the many identical copies distributed over the genome, but a small subset of these which had mutated away from the master sequence in the same manner.

On this picture, each set of repeated sequences might be 'tested' from time to time in evolution by the production of a control macromolecule (for example, a special protein) to recognize those sequences. If this produced a favourable result, natural selection would confirm and extend the new mechanism. If not, it would be selected against and discarded. Such a process implies that most sets of repreated sequences will never be of use since, on statistical grounds, their members will usually be in unsuitable places.

It thus seems unlikely that all selfish DNA has acquired a special function, especially in those organisms with very high C values. Nor do we feel that if one example of a particular sequence acquires a function, all the copies of that sequence will necessarily do so. As selfish DNA is likely to be distributed over the chromosomes in rather a random manner, it seems unlikely that every copy of a potentially useful sequence will be in the right position to function correctly. For example, if a specific sequence within an intron were used to control the act of splicing that intron, a similar sequence in an untranscribed region between genes would obviously not be able to act in this way.

In some circumstances, the sheer bulk of selfish DNA may be used by the organism for its own purpose. That is, the selfish DNA may acquire a nonspecific function which gives the organism a selective advantage. This is the point of view favoured by Cavalier-Smith in a very detailed and suggestive article[12] which the reader should consult. He proposes that excess DNA may be the mechanism the cell uses to slow up development or to make bigger cells. However, we suspect that both slow growth and large cell size could be evolved just as well by other more direct mechanisms. We prefer to think that the organism has tolerated selfish DNA which has arisen because of the latter's own selective pressure.

Thus, some selfish DNA may acquire a useful function and confer a selective advantage on the organism. Using the analogy of parasitism, slightly harmful infestation may ultimately be transformed into a symbiosis. What we would stress is that not all selfish DNA is likely to become useful. Much of it may have no specific function at all. It would be folly in such cases to hunt obsessively for one. To continue our analogy, it is difficult to accept the idea that all human parasites have been selected by human beings for their own advantage.

Life style

The effect of nonspecific DNA on the life style of the organism has been considered by several authors, in particular by Cavalier-Smith[12] and by Hindergardner[8]. We shall not attempt to review all their ideas here but instead will give one example to show the type of argument used.

Bennett[22] has brought together the measurements of DNA content for higher herbaceous plants. There is a striking connection between DNA content per cell and the minimum generation time of the plant. In brief, if such an angiosperm has more than 10 pg of DNA per cell, it is unlikely to be an ephemeral (that is, a plant with a short generation time). If it is a diploid and has more than 30 pg of DNA, it is highly likely to be an obligate perennial, rather than an annual or an ephemeral. The converse, however, is not true, there being a fair number of perennials with a DNA content of less than 30 pg and a few with less than 10 pg. A clear picture emerges that if a herbaceous plant has too much DNA it cannot have a short generation time.

This is explained by assuming that the extra DNA needs a bigger nucleus to hold it and that this increases both the size of the cell and the duration of meiosis and generally slows up the development of the plant. An interesting exception is that the duration of meiosis, is, if anything, shorter for polyploid species than for their diploid ancestors[23]. This suggests that it is the ratio of good DNA to junk DNA rather than the total DNA content which influences the duration of meiosis.

An analogous situation may obtain in certain American species of salamander. These often differ considerably in the rapidity of their development and of their life cycles, the tropical species tending to take longer than the more temperate ones. Drs David Wake and Herbert MacGregor (personal communication) tell us that preliminary evidence suggests that species with the longer developmental times often have the higher *C* values. This appears to parallel the situation just described for the herbacious plants. It remains to be seen if further evidence will continue to support this generalization. (See the interesting paper by Oeldorfe *et al.*[25] on 25 species of frogs. They conclude that 'genome size sets a limit beyond which development cannot be accelerated'.)

Testing the theory

The theory of selfish DNA is not so vague that it cannot be tested. We can think of three general ways to do this. In the first place, it is important to know where DNA sequences occur which appear to have little obvious function, whether they are associated with flanking or other sequences of any special sort and how homologous sequences differ in different organisms and in different species, either in sequence or in position on the chromosome. For example, it has recently been shown by Young[24] that certain intermediate repetitive sequences in *Drosophila* are often in different chromosomal positions in different strains of the same species.

Second, if the increase of selfish DNA and its movement around the chromosome are not rare events in evolution, it may be feasible to study, in laboratory experiments, the actual molecular mechanisms involved in these processes.

Third, one would hope that a careful study of all the nonspecific effects of extra DNA would give us a better idea of how it affected different aspects of cellular behaviour. In particular, it is important to discover whether the addition of nonspecific DNA does, in fact, slow down cells metabolically and for what reasons. Such information, together with a careful study of the physiology and life style of related organisms with dissimilar amounts of DNA, should eventually make it possible to explain these differences in a convincing way.

Conclusion

Although it is an old idea that much DNA in higher organisms has no specific function[8–12], and although it has been suggested before that this nonspecific DNA may rise to levels which are acceptable or even advantageous to an organism[8,12], depending on certain features of its life style, we feel that to regard much of this nonspecific DNA as selfish DNA is genuinely different from most earlier proposals. Such a point of view is especially useful in thinking about the dynamic aspects of nonspecific DNA. It directs attention to the mechanisms involved in the spread and evolution of such DNA and it cautions one against looking for a special function for every piece of DNA which drifts rapidly in sequence or in position on the genome.

While proper care should be exercised both in labelling as selfish DNA every piece of DNA whose function is not immediately apparent and in invoking plausible but unproven hypotheses concerning the details of natural selection, the idea seems a useful one to bear in mind when exploring the complexities of the genomes of higher organisms. It could well make sense of many of the puzzles and paradoxes which have arisen over the last 10 or 15 years. The main facts are, at first sight, so odd that only a somewhat unconventional idea is likely to explain them.

We thank W. Ford Doolittle and C. Sapienza for showing us their article before publication, and Drs D. Wake and H. Macgregor for allowing us to quote some of their unpublished conclusions about salamanders. This work was supported by the Eugene and Estelle Ferhauf Foundation, the J. W. Kieckhefer Foundation, the Ahmanson Foundation and the Samuel Roberts Noble Foundation.

1. Dawkins, R. *The Selfish Gene* (Oxford University Press, 1976).
2. Walker, P. M. B. in *Human Genetics: Possibilities and Realities*, 25–38 (Excepta Medica, Amsterdam, 1979).
3. Crick, F. H. C. in *From Gene to Protein: Information Transfer in Normal and Abnormal Cells* (eds Russell, T. R., Brew, K., Faber, H. & Schultz, J.) 1–13 (Academic, New York, 1979).
4. Crick, F. H. C. *Science* **204**, 264–271 (1979).
5. Doolittle, W. F. & Sapienza, C. *Nature* **284**, 601–603 (1980).
6. Callan, H. G. *J. Cell Sci.* **2**, 1–7 (1967).
7. Thomas, C. A. *A. Rev. Genet.* **5**, 237–256 (1971).
8. Hinegardner, R. in *Molecular Evolution* (ed. Ayala, F. J.) 179–199 (Sinauer, Sunderland, 1976).
9. Commoner, B. *Nature* **202**, 960–968 (1964).
10. Ohno, S. *J. hum. Evolut.* **1**, 651–662 (1972).
11. Comings, D. E. *Adv. hum. Genet.* **3**, 237–436 (1972).
12. Cavalier-Smith, T. *J. Cell Sci.* **34**, 274–278 (1978).
13. Galan, G. A., Chamberlin, M. E., Hough, B. R., Britten, R. J. & Davidson, E. H. in *Molecular Evolution* (ed. Ayala, F. J.) 200–224 (Sinauer, Sunderland, 1976).
14. Bishop, J. O. *Cell* **2**, 81–86 (1974).
15. Lewin, B. *Cell* **4**, 11–20 (1975); *Cell* **4**, 77–93 (1975).
16. Lomedico, P. *et al. Cell* **18**, 545–558 (1979).
17. Bernards, R., Little, P. F. R., Annison, G., Williamson, R. & Flavell, R. A. *Proc. natn. Acad. Sci. U.S.A.* **76**, 4827–4831 (1979).
18. Crick, F. H. C. *Eur. J. Biochem.* **83**, 1–3 (1978).
19. Britten, R. J. & Davidson, E. H. *Science* **165**, 349–358 (1969).
20. Davidson, E. U. & Britten, R. J. *Science* **204**, 1052–1059 (1979).
21. Wilson, A. C. in *Molecular Evolution* (ed. Ayala, F. J.) 225–236 (Sinauer, Sunderland, 1976).
22. Bennett, M. D. *Proc. R. Soc.* **B181**, 109–135 (1972).
23. Bennett, M. D. & Smith, J. B. *Proc. R. Soc.* **B181**, 81–107 (1972).
24. Young, M. W. *Proc. natn. Acad. Sci. U.S.A.* **76**, 6274–6278 (1979).
25. Oeldorfe, E., Nishioka, M. & Bachmann, K. *Sonderdr. Z.F. Zool., System Evolut.* **16**, 216–24 (1978).