Selfish DNA

In two review articles in the 17 April issue Doolittle and Sapienza (p. 601) and Orgel and Crick (p. 604) separately suggested that much of the DNA in the genome of higher organisms could be described as 'selfish'. They argued that such DNA has no appreciable phenotypic effect and functions only to ensure its own self-preservation within the genome. This view point stimulated a great deal of comment, some of which was published in the issue of 26 June (p. 617). Now the original authors have joined up with one of their critics and reassessed their ideas in the two articles below. A further comment is added by H. K. Jain.

from L.E. Orgel, F.H.C. Crick and C. Sapienza

DIFFICULTIES have been caused by the words 'selfish', 'junk', 'specific' and 'phenotype' that were used in the two reviews of selfish DNA[1,2].

Many people dislike the term 'selfish DNA' and a more acceptable alternative might be 'parasitic DNA'. The word 'parasitic' does not imply that the DNA can move between individuals, though certain viral DNAs might do this. It does imply that such DNA can usually move between different chromosomes in the same cell.

The word 'junk' also seems to arouse strong feelings. The idea behind it can be clarified by considering what is meant by 'specific'. We consider a sequence highly specific if the change of any one of its bases almost always has a considerable effect on the organism. An example would be the recognition site for a physiologically relevant restriction enzyme. At the other extreme are sequences whose deletion or extensive alteration would produce a negligible effect. Such sequences could reasonably be called junk. However, there is probably a continuum between these two extremes, including fairly specific sequences, where the alteration of most bases will produce some effect (many sequences coding for protein, and the different signals for starting and stopping transcription are likely to be of this type) and sequences whose deletion or extensive alteration will usually produce a small effect, such as a change in the local rate of recombination. In some cases close similarity of two sequences may be important, while the base sequences themselves may matter hardly at all — for example, within the introns of two neighbouring versions of a gene. The word 'junk' is perhaps too broad to cover all those cases for which the effect of sequence on the phenotype of an organism is small or zero. We hope a more precise terminology will evolve as the facts become better known.

The word 'phenotype' has also caused difficulties in spite of Doolittle and Sapienza's careful use of 'organismal phenotype' to make their meaning clear. We obviously need two words: one to refer to the phenotype of the organism and the other to apply solely to the 'phenotype' of the parasitic DNA, a distinction we would certainly make in the case of a true parasite. For the former we would suggest 'organismal phenotype' and for the latter, following Cavalier-Smith[3], 'intragenomic phenotype', but we would allow the word 'phenotype' alone to be used when the context makes the meaning clear.

In our original definition we said that selfish DNA had 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. By 'phenotype' we meant organismal phenotype. We intended 'specific' to be understood as 'highly specific' or 'fairly specific' in the discussion above. However it has been pointed out to us by R. Pritchard[4] that 'no... contribution' is unnecessarily strict. It would have been more useful to include also DNA which made a small contribution to the organismal phenotype, either positive or negative. An example of the latter might be a viral DNA which became part of the genome.

There is obviously a continuum of possible selective advantages (positive or negative) to the organism. We had excluded from our definition of selfish DNA those cases where the selective advantage is very high. To decide whether a repeated sequence is parasitic or not, one must determine whether the presence of the repeated sequence in the population is mainly due to the efficiency with which the sequence spreads intragenomically or mainly due to the reproductive success of those individuals in the population who possess repeated copies of the sequence. Only in the former case do we consider it useful to use the term selfish or parasitic DNA, as opposed to useful or symbiotic DNA — the borderline between the two may not be sharp.

In considering the spread of parasitic DNA one should not underestimate the power of natural selection. For example, if a particular transposon was inserted at random, it would run the risk of inactivating many genes and thus be selected against. A transposon which usually inserted at sites between genes would be at a selective advantage. Sites very near essential genes (as pointed out by Bruce Grant[5]) may be harder to delete than those in the middle of long stretches of junk and so parasitic DNA in the former positions is likely to survive longer. Effects of this type would lead to the selection of selfish DNA sequences that inserted preferentially at special sites in the genome.

Competing theories differ in their analysis of the factors determining the amount of non-specific DNA and of the way in which it comes into existence. Although we cannot at present decide on the quantitative contribution of the different types of non-specific DNA to the genome, it is still helpful to classify the various theories.

We proposed[2] that the amount of non-specific DNA present in a given genome is often determined by the balance between the intragenomic spreading of selfish sequences and phenotypic selection against excess DNA — the weaker the phenotypic selection against non-specific DNA the larger the DNA content of the genome. In another group of theories it is proposed that there is an optimal DNA content for each organism, which may be substantially greater than the amount of specific DNA that is needed to define the phenotype. The amount of non-specific DNA is then principally determined by the difference between the optimal DNA content and the essential content of specific DNA. The theories are not mutually exclusive, but differ substantially in emphasis in their explanation of C-values.

Cavalier-Smith's proposal[3,6] is an interesting example of an 'optimal DNA content' theory. One of his ideas, which we misinterpreted in our previous paper[2], is that in large cells, particularly in oocytes, the transport of messenger RNA across the nuclear membrane may become a limiting factor and that the only way to increase the

L.E. Orgel and F.H.C. Crick are at The Salk Institute, San Diego, California and C. Sapienza is in the Department of Biochemistry, Dalhousie University, Halifax, Canada.

The "selfish DNA" design at the top of the page was created by Linda Angeloff-Sapienza of Halifax, Canada and originally appeared on the cover of the issue of 17 April, 1980.

rate of transport is by increasing the number of nuclear pores by extending the surface of the membrane. If the area of nuclear membrane is determined by the DNA content of the nucleus, it follows that selection for a larger cell must lead to an increase in the DNA content of the genome. Thus, rather surprisingly, extra non-specific DNA is selected for because it allows such a cell to grow *faster*. While we do not question the logic of the argument, given the various assumptions, we do not find all the assumptions particularly plausible. It may be that there is sometimes selection for increased cell volume and increased nuclear volume. In cells so selected, non-specific DNA can accumulate. Whether it does so because large cells with large nuclei require such accumulation, or because they simply permit it remains to be seen. We feel that more experimental work is needed to unravel the complexities of the situation. In particular, we should like to know in which stages and in which organisms the surface of the nuclear membrane is saturated with nuclear pores.

Cavalier-Smith[3] also cites the widely different DNA contents of germ cells and somatic cells in some invertebrates as evidence against the selfish DNA hypothesis. However these observations can also be explained in terms of the selfish DNA theory. Such DNA 'needs' only to remain in the germ line to function parasitically. On the other hand, organismal selection might sometimes be stronger against surplus DNA in the soma than in the germ line. Thus representation in the germ line but not in the soma may sometimes be an optimal strategy for parasitic DNA. As for B chromosomes, in many cases the evidence appears to us to give some support to the idea (originally proposed by Östergren[7] in 1945) that they are largely parasitic, but there is certainly evidence that they sometimes have phenotypic effects which may possibly be useful[8,9].

Smith[10] has pointed out that the DNA of vertebrates usually has about 42 per cent GC whereas the GC content of invertebrate and prokaryotic DNA varies over a much wider range. The theory of parasitic DNA has rather little to say on this point. There are many factors which might affect the GC content of an organism's DNA. If much of the parasitic DNA has descended rather recently from insertion elements which themselves originally coded for proteins, then it would not be surprising if their present GC content were similar to that of genes which still code for protein. This may, perhaps, explain the constancy of GC in vertebrates.

As for our own ideas, we now feel that there may perhaps be reasons why too *little* DNA can in some cases produce a selective disadvantage. For example, Zuckerkandel[11] has suggested that there may be a minimum size for a 'domain' necessary for stability of the chromatin in the folded state. Thus a domain containing only a few genes might benefit from having some nonspecific DNA as 'padding'. This would mean that there is indeed an optimal amount for total DNA.

In our original paper[2] we feel that we did not put enough emphasis on the distinction between sequences which are repeated, exactly or nearly exactly, in many tandem repetitions and sequences which are more widely dispersed over the chromosomes and which occur in only one or a few copies in any one place. It seems plausible that these two types of sequence evolved different mechanisms. It is possible that the mechanisms generating the tandemly repeated type are usually more 'ignorant' (in Dover's sense[12]) than the more dispersed type. If the latter have any specific function it is likely to be that of the control, at one level or another, of gene expression, whereas the tandemly repeated type seem more likely to influence chromosome mechanics.

One possibility to which we feel we should have given more weight is that of 'dead genes', also called 'pseudogenes'[13,14]; that is, sequences which can no longer code for a protein (or a structural RNA) but which appear to have descended from a sequence that did. Whether these conform to our definition of parasitic DNA remains to be seen, but we suspect this is unlikely, since they usually exist in only a single copy, or as multiple tandem copies in only one place.

In our recent experience most people will agree, after discussion, that ignorant DNA, parasitic DNA, symbiotic DNA (that is, parasitic DNA which has become useful to the organism) and 'dead' DNA of one sort or another are all likely to be present in the chromosomes of higher organisms. Where people differ is in their estimates of the relative amounts. We feel that this can only be decided by experiment. We expect that due to the recent advances in genetic engineering and related techniques much sequence information will accrue in the near future. This should help to decide between the different alternatives. □

1. Doolittle, W. F. & Sapienza, C. *Nature* **284**, 601 (1980).
2. Orgel, L. E. & Crick, F. H. C. *Nature* **284**, 604 (1980).
3. Cavalier-Smith, T. *Nature* **285**, 617 (1980).
4. Pritchard, R. (personal communication).
5. Grant, B. (personal communication).
6. Cavalier-Smith, T. *J. Cell. Sci.* **34**, 247 (1978).
7. Östergren, G. *Bot. Notiser* **2**, 157 (1945).
8. Jones, R. N. *Int. Rev. Cytol.* **40**, 1 (1975).
9. Ames, A. & Dover, G. *Chromosoma* (in the press).
10. Smith, T. F. *Nature* **285**, 620 (1980).
11. Zuckerkandel, E. (personal communication).
12. Dover, G. *Nature* **285**, 618 (1980).
13. Loomis, W. *Devl Biol.* **30**, F3-F4 (1973).
14. See, for example, Proudfoot, N. *Nature* **286**, 840 (1980).