PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Introductory Remarks: DNA and Genes

P. M. B. Walker

Phil. Trans. R. Soc. Lond. B 1978 **283**, 305-307
doi: 10.1098/rstb.1978.0029

Phil. Trans. R. Soc. Lond. B. **283**, 305–307 (1978) [305]
Printed in Great Britain

II. DISTRIBUTION OF DNA BASE SEQUENCES

Introductory remarks: DNA and genes

By P. M. B. Walker
M.R.C. Mammalian Genome Unit, Department of Zoology, The University, Edinburgh EH9 3JT, U.K.

After the genetic code was discovered in the early 1960s, it was generally accepted that nearly all DNA in higher organisms was used to specify messenger RNA molecules at some time during their development. A small fraction could be set aside for the ribosomal and transfer RNAs and there was a problem about the rapidly turning over nuclear RNA which did not appear in the cytoplasm as message. By and large we considered that most DNA was potentially coding and the lone voices who talked of other kinds of DNA on the basis of somewhat flimsy evidence were largely ignored.

Recently these assumptions have been widely questioned and it is interesting to summarize why this should be so. A currently held, but perhaps somewhat extreme view would have it that in mammals there may be only 50 000 different proteins each coded by one or a few DNA sequences, that there were in addition regulatory sequences only long enough to be recognized by the proteins which bind to them, but that as much as 99 % of the DNA did not have a base-sequence rigidly conserved in evolution and was concerned with chromatin and chromosome structure and mitotic and meiotic functions.

Two main sources of evidence have contributed to this change of view. The first stems from genetic experiments concerned with mutation rates and the detailed genetic mapping of small regions of the *Drosophila* chromosomes. While the equation of one complementation group with one salivary chromosomes band, as earlier considered by Judd (Judd & Young 1974), is probably too simple a view, there is no genetic evidence to suggest that all the DNA in most bands specifies different proteins and would therefore be identifiable as 50–100 complementation groups in each band.

The other main strand of evidence comes from direct study of the DNA and its products. It is now well documented that there is a class of highly repetitive DNA, in which a short sequence of less than 20 bases is repeated many millions of times and which can comprise anything from a small percentage to 60 % of the mammalian genome. This satellite DNA does not code for proteins and is primarily located in the heterochromatic regions of chromosomes. It is often very variable in quantity and in sequence as between even closely related species. This variable amount of DNA is additional to, and does not substitute for, the euchromatin, where the functional genes are mostly located. Further, it is known that repetitive genes like those specifying ribosomal and transfer RNAs or coding for histones have spacer regions between the functional sequences. Messenger RNAs also have untranslated regions at both ends. These regions vary in size and in sequence between related species, thus sharing with satellite DNA the property of variability. They may also contain short conserved regions which could have a functional significance. In addition, most but not all messenger RNA molecules are derived from a precursor of high molecular mass, which can be several times

larger than the messenger sequence and which is the primary transcription product of the chromosomes. While satellite DNA can be shown to be additional to the normal complement of DNA, spacer and precursor sequences are clearly part of this complement.

The extrapolation of these results to the whole genome remains weak because although spacer sequences are often several times the length of the functional sequence, there is no evidence yet that all coding DNA has either untranscribed or untranslated regions 10–100 times larger attached to it. More fundamentally, no description of the genome in these terms will be acceptable until a function for this extra DNA is demonstrated. Functions, which may be different and unrelated, are needed both for satellite and spacer DNA, and many have been proposed. The two most popular current explanations for satellite DNA are that it is concerned with the recognition of homologues in meiosis, based on the unique patterns made by the several *Drosophila* satellite sequences in the different chromosomes (Goldring, Brutlag & Peacocke 1975; see also Brutlag, Appels, Dennis & Peacocke 1977) and that it is concerned with the regulation of recombination. There is no doubt that the presence of blocks of heterochromatin modifies the pattern of crossing-over in certain species (see, for example, Miklos & Nankivell 1976), but it does not follow that this is a sufficient selective force to explain the maintenance and amplification of satellite DNA. The problem is that we tend to expect that a defined function for DNA should have a preferred amount and kind of sequence for its optimal execution. On the other hand, our expectations may be entirely misleading. Spacers and satellites seem to be exceedingly variable and they occur widely.

The presence of non-coding DNA could be explained by a need to arrange the functional sequences in relation to higher order chromatin structures, and indeed chromosomes themselves may need to be of a minimum size if non-disjunction is to be consistently avoided. Evidence from lower eukaryotes which regularly shed a large part of their DNA, for example, the ciliate protozoon *Oxytricha* (Prescott & Murti 1974), strongly supports the view that here extra DNA is only needed for chromosomes which undergo meiosis and mitosis. The organization of chromosomes like the mammalian Y which have few genetic markers, but which have a large proportion of highly repetitive DNA (Cooke 1976, and this symposium) also supports the view that extra DNA is needed to make a stable chromosome.

My own view is that it is likely that chromosome structures have imposed the necessity for DNA with a mechanical or 'housekeeping' function over and above that required for the regulation and expression of the functional elements, which are under much stricter selective pressure than the intercalated spacer-like sequences or the satellites. In addition, the evidence suggests that there are many ways of solving these housekeeping problems, each species and even each individual may have a different optimal combination of solutions. These simply reflect the chance recombinational events which have effected these less 'essential' sequences, and which may themselves have properties like reduplication which facilitate these kinds of changes.

Another view is to retain a relatively small number of different protein products but to have longer regulatory sequences, which could be the middle-repetitive DNA discussed by Britten and Davidson. This DNA has interesting properties from this point of view. It is found scattered in the genome often next to sequences of about the length needed to code for proteins. It varies in length in different organisms but with many examples clustered around 300 base pairs and within an organism the sequences which comprise a related family show considerable differences. A hypothesis on regulatory effectors in terms of RNA, in particular the high molecular mass

nuclear RNA which does not leave the nucleus, is proposed by Davidson, Klein & Britten (1977). This still assigns a spacer function to the single copy sequence found transcribed in this kind of RNA but provides a regulatory rôle for the interspersed repetitive sequences. By contrast, proteins, if these are the regulatory effectors, only need a small number of bases for a unique recognition site and permit the pattern of middle repetitive sequences to be more the chance product of evolutionary processes, much like those we have come to expect from satellite and defined spacers. Regulatory sequences could then remain small, probably highly conserved and difficult to find because they are embedded in a matrix of sequences which are free to evolve more rapidly.

Fortunately it looks probable that recombinant DNA techniques will soon begin to provide the answer to the nature, and eventually the function, of the DNA surrounding the coding coding sequences (see, for example, Rubin, Finnegan & Hogness 1976).

References (Walker)

Cooke, H. 1976 *Nature, Lond.* **262**, 182.
Brutlag, D. L., Appels, R., Dennis, E. S. & Peacocke, W. J. 1977 *J. molec. Biol.* (In the press.)
Davidson, E. H., Klein, W. H. & Britten, R. J. 1977 *Devl Biol.* **55**, 69.
Goldring, E. S., Brutlag, D. L. & Peacocke, W. J. 1975 In *The eukaryote chromosome* (eds W. J. Peacocke R. D. Brock), p. 47. Canberra: Australian National University Press.
Judd, B. H. & Young, M. W. 1974 *Cold Spring Harb. Symp. quant. Biol.* **38**, 573.
Miklos, G. L. & Nankivell, R. N. 1976 *Chromosoma* **56**, 143.
Prescott, D. M. & Murti, K. G. 1974 *Cold Spring Harb. Symp. quant. Biol.* **38**, 609.
Rubin, G. M., Finnegan, D. J. & Hogness, D. S. 1976 *Progr. nucl. Acid Res. molec. Biol.* **19**, 221.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES