

# General Model for the Chromosomes of Higher Organisms

FRANCIS CRICK

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH

The model suggests that chromosomal DNA falls into two classes: globular DNA (containing unpaired regions for control) and a much smaller fraction consisting of fibrous DNA which alone codes for proteins.

I WISH to propose a general model for the structure of the chromosomes of higher organisms\*. This model is derived from ideas and data from many sources†. Because I have found it impossible to set out my ideas and the supporting evidence in a short space, I merely summarize here my conclusions. A much fuller account is in preparation and will be submitted for publication in the near future.

The model assumes that the DNA in a chromatid is a very long mononeme (see the review by Prescott<sup>2</sup> and the recent careful work by Laird<sup>3</sup>), which probably runs continuously from one end of the chromatid to the other.

\* I have used the term higher organisms rather than eukaryotes because I want to avoid having to discuss, at this stage, the chromosomes of various lower eukaryotes such as the dinoflagellates and the fungi.

† Proper acknowledgments will be given in the fuller paper, but I cannot refrain from mentioning here the very stimulating theoretical paper by R. J. Britten and E. H. Davidson<sup>1</sup>, which the reader is strongly recommended to read in parallel with this one. It contains extensive references to the earlier literature.

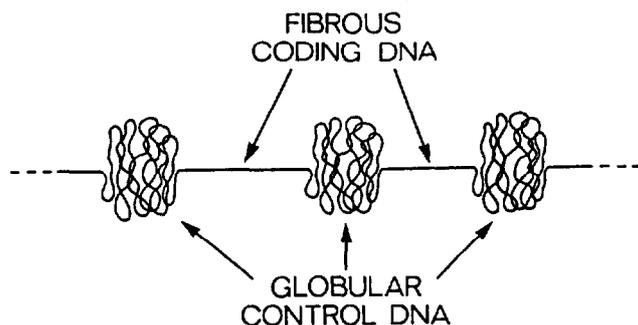
The model has three basic features. (1) The coding sequences of the DNA (that is, all those sequences which code for polypeptide chains) are postulated to be mainly, if not entirely, in the interbands (as visualized in the giant polytene chromosomes of the Diptera<sup>4</sup>). The bands, which contain all but a few per cent of the DNA, are identified as the control elements. This is illustrated diagrammatically in Fig. 1. A genetic complementation group is usually contained in a band plus an interband.

Thus on this view most of the DNA in higher organisms does not code for proteins but is used for control purposes, as already suggested by F. Vogel<sup>5</sup> in 1964. I have calculated that the average‡ amount of DNA (per mononeme) in an interband of *Drosophila*<sup>7,8</sup> is enough to code for an "average" protein of molecular weight 30,000–40,000. I have, therefore, adopted this speculation as a good working hypothesis.

(2) The central idea is that the recognition sites, needed for control purposes in higher organisms, are mainly unpaired single stranded stretches of double stranded DNA. I call this the Unpairing Postulate and it is set out diagrammatically in Fig. 2. It has been derived from a theoretical consideration of the general nature of protein molecules and the probable variety and length of the base sequences which need to be recognized in higher organisms. A particular type of example of this postulate has already been put forward by Gierer<sup>9</sup>.

The argument I give is a general one. It springs from the

‡ The average amount of DNA in the mononeme of an interband had been estimated previously by Beermann<sup>4</sup>.



**Fig. 1** An extremely schematic drawing of the proposed general structure of the DNA of the chromatid. The line represents part of the continuous DNA molecule in the monomeric chromatid. The straight portions correspond to the interband regions of the giant polytene chromosomes of the Diptera, which are postulated to be similar in their general character to the corresponding interphase chromatids, which are the active form. The mitotic chromosome is relatively inert<sup>5</sup>. The DNA sequences coding for protein are postulated to be mainly, if not entirely, in these extended regions. For convenience this DNA is referred to as fibrous DNA. The intricately folded regions correspond to the bands seen in the polytene chromosome<sup>4,7</sup>. No attempt has been made to represent their detailed structure. They are postulated to be the sites of the control regions. The model implies that a genetic complementation group is usually contained in either an interband plus a band or an interband plus part of the bands on either side. When a gene is active the bands are probably at least partly unfolded<sup>6</sup>. The globular DNA is certainly complexed with chromosomal proteins<sup>10</sup>, the fibrous DNA probably so. Thus both should be more strictly referred to as nucleoprotein.

fact that for all proteins whose tertiary structure is known, the active site is, to a first approximation, a shallow groove or a cavity and not a protruding piece of the protein structure. The former structures can rather easily be made to provide highly specific interactions, both with small molecules or with extended polymers. However, in double-helical nucleic acid the specific groups on the bases are not protruding but are themselves in one or other of the two grooves formed by the phosphate-sugar backbones. Thus, it might not be easy to design a protein of reasonable size to recognize more than a limited number of base pairs, especially when one remembers the twisted nature of the normal double helix. The postulate has even more force if single stranded RNA is also used to recognize the control sequences on the DNA, as favoured (with reservations) by Britten and Davidson<sup>1</sup>.

My argument is that when very large numbers of different sequences need to be recognized (which implies that the sequences must not be too short), it will pay to unwind the double helix before recognition. This may be expensive to arrange but in the long run it will provide a much greater abundance of versatility.

(3) The forces and energy needed to unpair the recognition stretches of the DNA are provided by the combination of the DNA with chromosomal proteins—probably the histones<sup>10</sup>. Although the three-dimensional configuration of a band may be very intricate, such globular structures may be based on structural motifs of various kinds. The most obvious one is a simple supercoiled DNA, that is, a helical double helix, as already suggested by various workers<sup>11,12</sup>, but other more complicated structures are possible<sup>11</sup>. One such example is illustrated in Fig. 3. As far as I know this suggestion is a novel one.

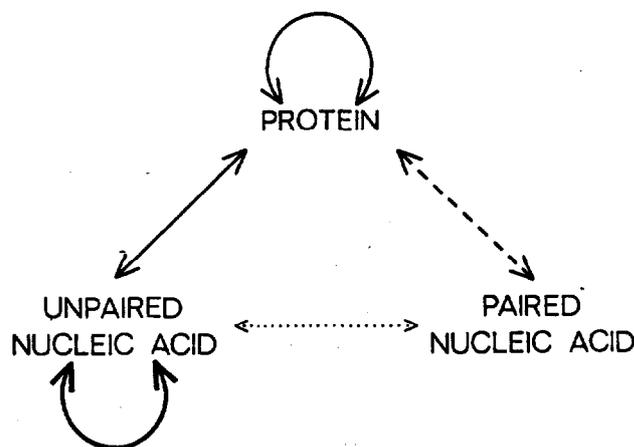
The general property of this family of structures is that there are lengths (probably of hundreds of base pairs) of DNA whose exact base sequence matters very little, interspersed with shorter stretches (perhaps of a hundred bases or so) of specific sequences which are probably repeated at very many places along the DNA. On this view the role of the histones is not merely to cover up the DNA but to help the DNA to expose itself in the right places.

Plausible general reasons can be given why the bands (the

control elements) are so large compared with the interbands (the coding elements). In the first place, the type of structural motif outlined in Fig. 3 cannot be constructed on too small a scale. Moreover, as Britten and Davidson<sup>1</sup> have pointed out in their Fig. 1A, multiple control elements may well be needed adjacent to each particular coding sequence. In addition, a set of similar elements may be required within certain bands to help provide a graded response. Finally, it appears to be a general rule that intricate three-dimensional biological structures are always bigger than one might naively expect. The examples of globular proteins, transfer RNA and ribosomes spring to mind.

My fuller paper will discuss possible mechanisms for the formation during evolution of these large control regions. It would seem likely that both tandem repetition and translocation will be involved. Whatever the origin of tandemly replicated sequences (satellite DNA<sup>13</sup> or otherwise), when they are first formed there will be an exact repeat of the sequences both in the paired region (such as the stem of Fig. 3) and in the regions which become unpaired (such as the loop of Fig. 3). However, during evolution, mutations will accumulate at different rates in these different regions. The paired regions will diverge rapidly, since the exact base sequences there do not matter appreciably, but changes in the unpaired regions will occur less rapidly, if at all, because they have to interact with other molecules during control operations. Thus, newly evolved bands (or parts of bands) would be expected to have a closer degree of tandem repetition of their base sequences than phylogenetically older ones.

The model, in its simplest form, suggests that the number of different proteins normally produced by higher organisms may not be much more than a few thousand for *Drosophila* and



**Fig. 2** Each line represents the possibility of a highly specific interaction between two macromolecules. One molecule is from the class at one end of the double arrow and the other is from the class at the other end. Notice that no distinction is made between RNA and DNA. Instead, a strong distinction is made between paired nucleic acid (meaning a stretch of double helix) and unpaired nucleic acid which can be either single stranded nucleic acid or unwound stretches of an originally double helical structure. The latter may or may not be refolded to some extent into three-dimensional structures having a complex mixture of paired and unpaired regions<sup>9</sup> such as is found in transfer RNA. The dotted line represents the formation of a triple helix, such as poly A+2 poly U. The dashed line represents the recognition by a specific protein molecule of a particular base sequence of base-paired double stranded nucleic acid. The solid lines represent interactions of abundant versatility. The very thick line emphasizes that for any sequence of the usual monomers (base sequence) there is always a complementary base sequence. The diagram does not deal with relatively unspecific interactions, such as the interaction between a particular protein and a DNA backbone independent of the latter's base sequence. The Unpairing Postulate states that the interactions used for control in higher organisms will mainly be chosen from those shown by solid lines in the figure. This implies that double helical DNA will usually have to be unwound at the recognition sites.

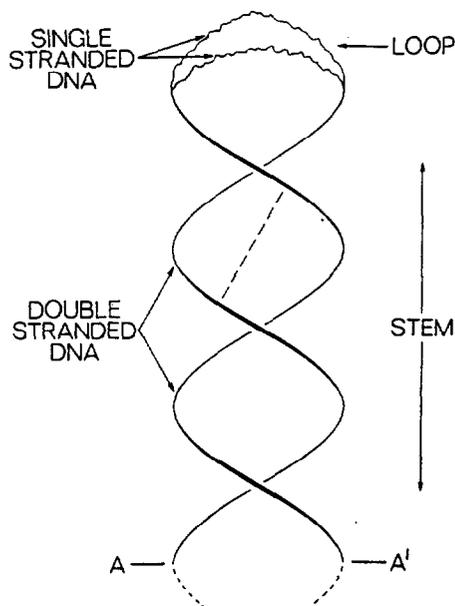


Fig. 3 An example of a possible structural motif within the globular DNA: a twisted hairpin constructed from part of a single length of double stranded DNA. The loop itself has become unpaired due to the untwisting effect produced by the stem. The DNA in the stem remains double stranded and forms a double helical double helix, stabilized by chromosomal proteins, probably mainly by histones. The figure is highly schematic and the details should not be taken too literally. For example, one type of histone (either as a monomer or a dimer) may bridge two adjacent helices in the sort of way shown by the dashed line. Another type of histone may do the same by lying horizontally, thus bridging strands on opposite sides of the axis of the structure. The loops themselves may or may not be stabilized in special configurations by chromosomal proteins or by folding back on themselves. Since the histones interact mainly with the backbone<sup>10</sup> of the DNA, with little respect for base sequence, the actual base sequence of the lengths of DNA in the stem is not important, at least to a first approximation. The loops are postulated to be the sites of the actual control elements. Any particular loop may have an unwinding sequence (to help localize the unwinding—possibly a sequence of A's on one strand), a promoter sequence for the RNA polymerase, or an operator sequence (whether for positive or negative control is left open) or, more likely, some combination of these sequences. The base sequences for the first two functions may be much the same in very many different loops. The operator sequences may or may not be repeated in various other loops, but in general such sequences are likely to be repeated rather less often than the other two types postulated. An occasional loop may contain an initiator sequence for DNA replication. Since the normal DNA double helix is right handed, the superhelix is more likely to be left handed (as shown in the figure) for mechanical reasons. I thank Drs Graeme Mitchison and David Baillie for instructing me on this point. Strictly speaking, either hand is possible for the superhelix if the histones can distort the basic twist of the DNA double helix in the appropriate way. If the second type of histone postulated above interacts with its neighbours (above and below, near the axis of the structure), this interaction could itself impose one particular hand of twist, since protein molecules are intrinsically handed. The structure shown is only one of a family of similar structures. The simple helical double helix is the first member of the family. Another example could be constructed by intertwining a pair of hairpins (of double helical DNA) to form a stem having a quadruply helical double helical structure, with four single stranded stretches in the loop region. Whether the single strands within a loop region can interact, by complementary base pairing, with similar stretches in other loops in the same "band" of a single chromatid (or also, in the second structure mentioned, within the same loop region) is an open question. I expect this to happen in the highly repetitive "satellite" sequences<sup>13</sup> found in the heterochromatin near the centromeres<sup>14,15</sup>. In the euchromatin the answer must, in part, depend on the relative location of different hairpins within the "band" of one chromatid. This is not easy to decide, although various attractive models are possible. (For example, a double-headed structure might be formed, roughly described by the operation of a dyad axis along the line AA' in the figure.) Be that as it may, the single stranded regions are postulated to mediate the highly specific lining up of the bands of the polytene chromosome<sup>7</sup> by forming complementary base-pairs between adjacent "chromatids" in the same band. A similar interaction may perhaps take place in meiosis.

some tens of thousands for man. It is not, however, completely excluded that, in special cases, multiple coding sequences may be hidden within some of the globular bands, in which case the numbers could be higher. The amount of these special bands, if they exist, might differ markedly in different kinds of higher organisms.

The model, which is logically coherent, appears to me to be compatible with a very large amount of experimental data obtained using very different techniques. These include rough estimates from genetic data<sup>16</sup> of the number of "genes" in *Drosophila* and man, the correlation between the number of bands plus-interbands and genetic complementation groups<sup>17</sup>, the specific pairing between the bands of the giant polytene chromosomes<sup>7</sup> shown by the study of inversions and so on, the nature<sup>13-15</sup> and general effects of the heterochromatin<sup>18</sup>, the large amount of data on nucleic acid hybridization<sup>1,19</sup> and the formation of circles by the technique of Thomas and his colleagues<sup>20</sup>.

It can also be accommodated to the data on the "puffing" of polytene chromosomes<sup>4</sup>, the general nature of the heterogeneous rapidly turning over nuclear RNA<sup>21-24</sup> and the apparent absence of polycistronic messengers in higher organisms. It is not incompatible with the very scanty X-ray studies of chromatin<sup>12,25</sup> and the electron microscope pictures and measurements<sup>11</sup>.

Although the model is speculative and not fully detailed, and raises at least as many questions as it attempts to answer, I hope it may serve as a focus for discussion and for the design of experiments.

In addition to thanking many colleagues for their patience in explaining their results and for discussing these problems with me, especially Sydney Brenner, Leslie Orgel, Peter Walker (see his remarks reported in ref. 26), Gordy Tomkins and David Baillie, I thank especially the organizers of the specialists' meeting in May of this year at Port Cros<sup>26</sup> for inviting me to attend.

*Note added in proof.* On the amount of informational DNA in higher organisms see T. Ohta and M. Kimura, *Nature*, 233, 118 (1971).

Received September 6, 1971.

- <sup>1</sup> Britten, R. J., and Davidson, E. H., *Science*, 165, 349 (1969).
- <sup>2</sup> Prescott, D. M., in *Advances in Cell Biology*, 1 (edit. by Prescott, D. M., Goldstein, L., and McConkey, E.), 57 (North Holland, Amsterdam, 1970).
- <sup>3</sup> Laird, C. D., *Chromosoma*, 32, 378 (1971).
- <sup>4</sup> Beermann, W., in *Cell Differentiation and Morphogenesis*, 24 (North-Holland, Amsterdam, 1966).
- <sup>5</sup> King, D. W., and Barnhisel, M. L., *J. Cell Biol.*, 33, 265 (1967).
- <sup>6</sup> Vogel, F., *Nature*, 201, 847 (1964).
- <sup>7</sup> Lindsley, D. L., and Grell, E. H., *Genetic Variation in Drosophila melanogaster* (Carnegie Inst. Washington, Publ. 627, 1968).
- <sup>8</sup> MacInnes, J. W., and Uretz, R. B., *Science*, 151, 689 (1966).
- <sup>9</sup> Gierer, A., *Nature*, 212, 1480 (1966).
- <sup>10</sup> *Histones and Nucleohistones* (edit. by Philips, D. M. P.) (Plenum, New York and London, 1971).
- <sup>11</sup> DuPraw, E. J., *DNA and Chromosomes* (Holt, Rinehart and Winston, 1970).
- <sup>12</sup> Pardon, J. F., Wilkins, M. H. F., and Richards, B. M., *Nature*, 215, 509 (1967).
- <sup>13</sup> Southern, E. M., *Nature*, 227, 794 (1970).
- <sup>14</sup> Jones, K. W., *Nature*, 225, 912 (1970).
- <sup>15</sup> Pardue, M. L., and Gall, H. G., *Science*, 168, 1356 (1970).
- <sup>16</sup> Muller, H. J., in *Heritage from Mendel* (edit. by Brink, R. A.), 419 (University of Wisconsin Press, Madison, 1967).
- <sup>17</sup> Shannon, M. P., Kaufman, T. C., and Judd, B. H., *Genetics*, 64, s58 (1970).
- <sup>18</sup> Baker, W. K., *Adv. Genetics*, 14, 133 (1968).
- <sup>19</sup> Sutton, W. D., and McCullum, M., *Nature*, 232, 83 (1971).
- <sup>20</sup> Thomas, jun., C. A., Hamkalo, B. A., Misra, D. N., and Lee, C. S., *J. Mol. Biol.*, 51, 621 (1970).
- <sup>21</sup> Scherrer, K., Spohr, G., Granboulan, N., Morel, C., Grosclaude, J., and Chezzi, C., *Cold Spring Harbor Symp. Quant. Biol.*, 35, 539 (1970) and other papers in the same volume.
- <sup>22</sup> Georgiev, G. P., *J. Theoret. Biol.*, 25, 473 (1969).
- <sup>23</sup> Ryskov, A. P., and Georgiev, G. P., *Febs Letters*, 8, 186 (1970).
- <sup>24</sup> Coutelle, C., Ryskov, A. P., and Georgiev, G. P., *Febs Letters*, 12, 21 (1970).
- <sup>25</sup> Luzzati, V., and Nicolaieff, A., *J. Mol. Biol.*, 7, 142 (1963).
- <sup>26</sup> *Nature New Biology*, 231, 68 (1971).