Repeated DNA: Molecular Genetics of Higher Organisms

The elucidation of the genetic code has resulted in years of research on isolated sequences of bacterial DNA. Now emphasis is increasingly given to the molecular biology of higher organisms. One feature of nonbacterial cells that is of great interest is DNA sequences that occur many times in such cells. Attempts to describe the organization of these repeated sequences have led to tentative explanations of their role in molecular genetics.

Investigators whose studies dealt with nonbacterial (eukaryotic) DNA soon came upon these anomalous repeated sequences. Unlike bacterial DNA, which is thought to consist of a linear array of unique sequences, the DNA of each eukaryotic cell contains many copies of some of its DNA sequences. Moreover, there is a large variation in the amount of DNA in the cells of some closely related organisms. For example, some amphibian species have 10 to 20 times as much DNA as do other, closely related species, and these differences are thought to stem from differences in the number of times that certain sequences are repeated.

Researchers would like to know whether the repeated sequences are genes (DNA sequences that code for proteins) and would like to describe the organization of these sequences. The organization of repeated sequences, they believe, is almost certainly related to the control of genes.

Recently, several investigators have proposed models of eukaryotic DNA that describe the organization and function of repeated sequences. The models differ as to whether the repeated sequences are genes, but the several models agree that the repeated and unique sequences of many organisms are interspersed in specific patterns—patterns that may be common to all eukaryotic DNA.

Repeated DNA sequences were first suggested by experiments in which the two strands of DNA molecules were separated (when the strands were heated) and then allowed to reassociate. Two single strands of DNA will reassociate to form a double strand when two complementary nucleotide sequences collide. If the single strands contain repeated nucleotide sequences, it is more likely that two complementary sequences will collide. The amount of DNA that reassociates in a given interval of time can thus be correlated with the number of repeated nucleotide sequences in that DNA.

Unique Sequences Code for Protein

Roy Britten, then at the Carnegie Institution of Washington in Washington, D.C., and his colleagues studied repetitive sequences with this technique of reassociation. They sheared DNA into short fragments and studied their reassociation rates as a function of fragment length. They showed that certain portions of the DNA reassociate very rapidly and apparently contain sequences that are repeated many times. Other DNA sequences are apparently unique ones, and some researchers hypothesize that these may code for proteins.

Biochemical studies of protein synthesis provide one approach to answering the question of whether the unique sequences code for proteins. Researchers can isolate messenger RNA (mRNA) from the cytoplasms of cells and determine what portion of the DNA was transcribed when that mRNA was formed. Since mRNA is a complementary copy of DNA, it will bind to that portion of the DNA from which it was transcribed. The rate at which the mRNA and single strands of DNA reassociate is a function of the number of copies of DNA nucleotide sequences that are complementary to the mRNA.

The studies of the reassociation of DNA with mRNA have involved measurements of the rate that mRNA which is specific for a particular cellular protein reassociates with DNA or the rate that a mixture of all of the mRNA in a collection of cells (total mRNA) reassociates with DNA. It is necessary to obtain large quantities of mRNA for reassociation experiments and to positively identify the mRNA as such. Thus researchers who study specific mRNA's have restricted their investigations to mRNA's for those proteins that are produced in large quantities by certain cells. For example, there have been studies of hemoglobin mRNA from mouse and duck immature red blood cells, silk fibroin mRNA from cells of the posterior gland of the silk moth *Bombyx*, and histone mRNA from cells from sea urchin embryos at cleavage stage. Studies of total mRNA have, so far, been limited to studies of sea urchin mRNA from cells at the late gastrula stage of development, of mouse L cell mRNA, and of mRNA from the cellular slime mold *Dictyostelium*.

The results of these reassociation experiments are consistent with the hypothesis that most proteins are coded by unique DNA sequences. One to three DNA sequences reassociate with duck hemoglobin mRNA and, with the exception of histones, one DNA sequence reassociates with each of the other specific mRNA's mentioned. The DNA sequences that reassociate with the total mRNA's studied are also, apparently, unique sequences. For example, Britten and Eric Davidson of the California Institute of Technology have found that at least 95 percent of the total mRNA in sea urchins at late gastrula reassociates with DNA as though it were transcribed from unique sequences.

There is an exception to these indications that proteins are coded by unique sequences. The sequences that code for histones are repeated from 400 to 1200 times, depending on the species, and these repetitive histone-coding sequences are clustered on the DNA. Some researchers believe that the repetition and clustering of histone genes represent a control mechanism whereby the cells can produce a great deal of histone mRNA at that stage of the cell cycle (the S stage) when histones are synthesized. They believe that other mechanisms control the synthesis of most proteins and that an understanding of these control mechanisms may lead to an understanding of the DNA sequences that do not code for proteins. In order to propose models of such control mechanisms, investigators have sought models of DNA sequence organization.

Much of the research on sequence organization has been directed toward

attempts to locate the position of repeated DNA sequences with respect to nonrepeated sequences. Because the repeated sequences are apparently present in all eukaryotic DNA and because there are so many copies of each repeated sequence, Britten and Davidson suggest that such sequences may be organized in specific patterns that are important in gene regulation.

Britten and Davidson have recently proposed an organizational scheme for the DNA's of the sea urchin and the toad (*Xenopus*). They believe, on the basis of an analysis of the rate that single strands of DNA fragments reassociate as a function of fragment length, that about 50 percent of sea urchin and *Xenopus* DNA's consists of closely interspersed repetitive and unique sequences. The repetitive sequences are composed of about 200 to 400 nucleotides and the unique sequences of about 650 to 900 nucleotides. The remainder of the DNA consists of long unique sequences (at least 4000 nucleotides) interspersed with short repetitive sequences and a region (about 25 percent of the DNA) about which little is known. Since the sequence patterns in *Xenopus* are quantitatively similar to those in the sea urchin, Britten and Davidson speculate that such patterns may be a general feature of animal chromosomes.

Independent evidence for the universality of sequence patterns was obtained by Charles Thomas and his colleagues at Harvard Medical School. They found that, in a large variety of organisms, repeated DNA sequences of a given type are apparently clustered in short DNA segments. About half of the total DNA is found in these segments, perhaps interspersed with nonrepeated DNA. Their results need not be incompatible with those of Britten and Davidson, but their experimental approach allows them to undertake a more detailed analysis of certain aspects of chromosome structure, as compared to that of Britten and Davidson. Thus, a detailed analysis of fruit fly (*Drosophila*) chromosomes has led them to propose an interesting and controversial model for chromosome structure in this organism.

Thomas developed a method to study chromosome structure which is based on the following argument. If DNA that contained densely clustered repeated sequences were randomly broken, then it would be likely that many fragments contain the same nucleotide sequence at both fragment ends. If the randomly broken fragments are partially degraded by either a 3′ exonuclease or a 5′ exonuclease (an enzyme that selectively removes nucleotides from the 3′ or 5′ end of each DNA strand), each fragment will have a single strand of DNA at either end but will otherwise remain double stranded. Those fragments that have the same nucleotide sequence at either end will now have complementary strands of single DNA chains at either end. Such fragments can form rings, and the rings can be seen in an electron microscope.

Rings Reveal DNA Organization

Thomas and his associates studied DNA from salmon, trout, salamanders, mice, calves, and fruit flies and found that DNA fragments from all of these organisms formed rings. In contrast, no rings formed when DNA fragments were made from DNA of those organisms (bacteria and some of their viruses) that do not have repeated DNA sequences. By varying the fragment sizes, Thomas and his colleagues were able to analyze the patterns of repeated and unique sequences in these DNA's. In particular, they were able to obtain detailed information about the *Drosophila* chromosome.

When DNA from *Drosophila* salivary glands was fragmented and partially degraded by an exonuclease, the efficiency of ring formation turned out to be a function of the fragment length. Those fragments shorter than 1.5 micrometers (about 4500 nucleotides) and those longer than 5 to 10 μm formed fewer rings; those that were about 2 μm formed rings most efficiently. In that very short fragments formed few rings, there should be a small probability of obtaining a short fragment with the same nucleotide sequence at each end. The short segments consist of up to 3000 nucleotides; the repeated sequences are shorter than 1500 nucleotides. Thus the repeated sequences are separated by some other sequences on the DNA. The fact that long fragments formed few rings indicates that repeated sequences are clustered in segments whose length, Thomas calculates, is about 5 μm. Thomas sees so many rings that he proposes that different repeated sequences must be close together on the DNA.

Because of these results relating to sequence organization, Thomas proposes that the *Drosophila* chromosome is composed of DNA segments 5 μm long with sequences in tandem repetition bracketed by equal amounts of nonrepeated sequences. The *Drosophila* chromosome has been studied by genetic means, and Thomas's structural model has some interesting correlations with the model of this chromosome proposed by those who studied *Drosophila* genetics.

The chromosomes of *Drosophila* salivary glands appear as a sequence of dark bands when viewed in a light microscope. Most of the DNA is in these bands (only 5 percent is in interband regions), and each band can be associated with one genetic function. According to Burke Judd of the University of Texas at Austin, is is possible that each band contains coding information for only one protein. Since each band contains enough DNA to code for 30 proteins only 1/30 of *Drosophila* DNA need consist of genes.

Judd and his associates studied *Drosophila* bands by producing mutations in the X chromosome of the organism. They found that a mutation that damages a particular function can be associated with the alteration of the nucleotide sequence of one particular band. Another researcher, George Lefevre of the California State University in Northridge, sought to ascertain whether a chromosome that is broken (by x-rays or other mutagens) at the site of a band results in mutant effect. He found that a large proportion (50 to 60 percent) of these breaks do not result in mutant effects. This does not contradict the idea that only 1/30 of the DNA in a band codes for protein, but, as Judd points out, his results and those of Lefevre are only consistent with, but do not prove, the one-gene one-band hypothesis. There are two major difficulties with these experiments. First, investigators can only associate genes with their function in the cell. Genes are defined as sequences that code for proteins. Thus if one band consists of many interacting genes, any mutation that destroys that interaction will destroy the function associated with the band. The second difficulty is that, when researchers observe chromosome breakage, they cannot locate by cytological methods the exact position of the break in the band. It is possible that these breaks occur preferentially in the nongene portion of the band and that a great deal of the DNA in the band consists of genes.

If the one-gene one-band hypothesis is correct, either the information for

one protein is repeated many times in a band or most of the DNA in a band does not code for protein. Thomas believes that the information for one protein could be repeated many times in a band but points out that his experiments cannot prove that protein-coding sequences are multiply represented. Thomas can form rings from DNA that was thought to consist of unique sequences when it was analyzed by reassociation techniques. Thus the proposal that the protein-coding sequences are repeated many times does not necessarily contradict the reassociation experiments that indicate that proteins are coded by unique sequences.

Thomas bases his belief—that the protein-coding information could be repeated many times in a band—on his experiments with rings. These experiments suggest more or less exact tandem repetitions of sequences within a band. He estimates that the number of chromosomal regions that can form rings—those regions containing many copies of a repetitive sequence—equals the number of bands in the *Drosophila* chromosome. He proposes that at least 50 percent of a band consists of tandemly repeated copies of a gene. A gene must consist of about 1000 nucleotides to code for a protein. Thomas calculates that the repetitive sequences in a band range from 600 to 6000 nucleotides.

The minimum length of repetitive sequences in a ring can be calculated from the ring's stability. The longer the repeated sequence is, the more bonds will be formed when the ring is made and the more stable the ring will be. By investigating the thermal stability of *Drosophila* DNA rings, Thomas concluded that the rings are closed by repeated sequences of at least 200 nucleotides.

Thomas's model of chromosome structure differs from the model of Britten and Davidson. Britten and Davidson believe that the repeated sequences consist of only 200 to 400 nucleotides and that it is the longer unique sequences that are the genes. This controversy has been analyzed by James Bonner and Jung-Rung Wu at the California Institute of Technology. They proposed a model of the *Drosophila* chromosome that agrees with the model of Britten and Davidson and yet is consistent with Thomas's data. In addition, Bonner believes that the organization of the rat chromosome may resemble that of the *Drosophila* chromosome.

Bonner and Wu studied the *Drosophila* chromosome by a technique that combines reassociation of single DNA strands with examination by electron microscopy. They were able to determine the lengths and distribution of repetitive DNA sequences. They determined the lengths of repetitive DNA sequences by breaking the DNA into 800-nucleotide fragments, separating the fragment strands, and allowing the strands to reassociate under conditions of temperature, time, and strand concentration that favor the reassociation of repetitive DNA. They examined the reassociated strands by electron microscopy and found that the repeated segments are short—100 to 150 nucleotides—and so could not code for proteins.

Bonner and Wu then used similar techniques to determine the distribution of the repeated sequences. They broke DNA into short fragments, separated the fragment strands, allowed them to reassociate under conditions that favor the reassociation of repeated sequences, and measured the distance (on an electron micrograph) between reassociated segments on the long DNA strands. Most repeated segments were separated by 750 nucleotides, although some were separated by sequences that were as long as 3000 nucleotides.

Basing their model on these results and those of Thomas, Bonner and Wu propose that each band of the *Drosophila* chromosome comprises 30 to 35 unique sequences that are separated by short repeated sequences, and that the repeated sequences in a band are all alike. Each of the unique sequences, they say, could code for a protein since the genetic association of one band with one function does not preclude the possibility that each function depends on many proteins.

Bonner and Wu were led to propose this model by the coincidence between the number of families of repetitive sequences (4500) and the number of bands on the *Drosophila* chromosome (between 3500 and 5000). This coincidence suggested to them that each band could be associated with one family of repetitive sequences.

As a test of their model, Bonner and Wu calculated the probability that rings would form from fragments of chromosomes having the structure that they proposed. Their plot of the probability of ring formation as a function of fragment length coincides with Thomas's plot of his data. Moreover, they found that Thomas's data on the thermal stability of *Drosophila* DNA rings is consistent with their hypothesis that rings are formed by repeated sequences consisting of 100 to 150 nucleotides.

Rat DNA Resembles Drosophila DNA

Bonner has used his technique to study the chromosomes of the rat and once again he found that the repeated sequences are short (about 100 to 200 nucleotides). Rat DNA differs from *Drosophila* DNA, however, in that there are usually two short repetitive sequences between each unique sequence. The unique sequences of *Drosophila* DNA are about 750 nucleotides; half of the unique sequences of rat DNA are between 500 and 2000 nucleotides and the remainder are as long as 16,000 nucleotides.

Bonner believes that his model of *Drosophila* DNA can be extended to rat DNA. He proposes that rat DNA is organized in regions in which the same repetitive sequence separates various unique sequences in a region and plans to test this hypothesis by forming rings with rat DNA. The maximum circumference of the rings is the length of a region characterized by a repetitive sequence. If he can verify his hypothesis that rat and *Drosophila* DNA are organized in such regions, Bonner will have demonstrated unprecedented similarities in the structure of chromosomes from two vastly unrelated species.

Bonner's theories of chromosome structure are still mainly speculative; but his results, along with those of Britten and Davidson and of Thomas indicate that the chromosomes of a large variety of unrelated species are similarly organized. Although there is still some question as to whether repeated sequences code for proteins, the repeated and unique sequences appear to be arranged according to principles which, if they are not simple, are at least not complicated so that they discourage study. Thus it seems likely that the organization and function of repeated sequences in eukaryotic chromosomes may be characterized in the near future.—GINA BARI KOLATA

Additional Reading

1. E. Davidson, B. R. Hough, C. Amenson, R. Britten, *J. Mol. Biol.* **77**, 1 (1973).
2. C. Thomas *et al.*, *Cold Spring Harbor Symp. Quart. Biol.*, in press.
3. J. Bonner and J.-R. Wu, *Proc. Nat. Acad. Sci. U.S.A.* **70**, 535 (1973).
4. J.-R. Wu, J. Hurn, J. Bonner, *J. Mol. Biol.* **64**, 211 (1972).