The Organization of the Eukaryotic Genome

Giorgio Bernardi

Our understanding of the organization of the prokaryotic genome and of the regulation of its expression can be considered as fairly satisfactory. In contrast, the organization of genes in the eukaryotic genome is one of the major open problems in today's molecular biology. This issue can hardly be overestimated, since it is of capital importance for understanding evolution, differentiation, and oncogenesis.

We will attempt here, first, to introduce the major questions concerning the organization of eukaryotic genomes; second, to summarize the picture emerging from the experimental approach which has been most widely used in recent years, namely kinetic studies of DNA renaturation; and, third, to present a new approach developed in our laboratory along with the results obtained so far.

INTRODUCTION

A comparison of prokaryotic and eukaryotic cells reveals that the major differences concern the size, the structure, and the organization of the genome.

Genome Size

The amount of (nuclear) DNA per eukaryotic (haploid) cell (genome size), is constant in each species and covers an extremely wide range of values. Some fungi have genome sizes practically

equal to those of bacteria, whereas some animals and plants have genome sizes 10,000 times as large. A closer look at available data indicates that wide variations of genome sizes are found within single phyla, within single genera, and even within interbreeding species. It is unlikely that these differences imply corresponding differences in the amount of genetic information. If these differences are, therefore, neglected, and only the minimum genome sizes found in each phylum or order are considered, a ratio of about 1,000 is found between the smallest genome size of prokaryotes and the largest genome size found in eukaryotes, namely the genome size of mammals; furthermore, a well-defined trend for the minimum genome size to increase with evolution is evident.

Genome Structure

The genetic material of eukaryotes is structured in a much more complex way than that of prokaryotes, even if the latter is far from being just a double helix of DNA suspended in the cell sap. Eukaryotic DNA is organized into nucleosomes--complexes of histones H2a, H2b, H_3, and H_4 wrapped around DNA. The chromatin fiber, with its nucleosomes and its acidic proteins, is folded, probably with the help of histone H1, into a chromatid. Before the chromatid is replicated, there is only one chromatid (unit chromatid) per chromosome, at least according to the widely accepted mononeme theory. This unit chromatid is thought to contain a single thread of the DNA double helix and to be continuous in each chromosome (at least in yeast and *Drosophila*). At mitotic metaphase, the chromosome is made up of two (sister) chromatids.

Chromatids are arranged into a series of chromomeres, which have the appearance of characteristic bands in polytene chromosomes, namely chromosomes formed by hundreds of chromatids originated by internal replication. Structurally, chromomeres correspond to local packings of chromatin; functionally, they have been classically

considered as units of mutation, replication, and transcription, corresponding to Mendelian genes. These concepts are being further studied at the present time. In any case, chromomeres contain much more DNA than required by a gene coding for an average protein.

Another major difference between prokaryotic and eukaryotic cells is the segregation, in eukaryotes, of parts of the genome into cytoplasmic organelles (mitochondria, chloroplasts). This DNA seems to be structured according to the prokaryotic rather than the eukaryotic model.

Genome Organizations

The prokaryotic genome is basically made up of a series of structural genes coding for messenger RNA, (and therefore polypeptide chains), ribosomal RNA, and transfer RNA. Promoters (sequences binding RNA polymerase) and operators (sequences binding regulatory proteins, such as repressors) are found at the beginning of each polycistronic transcription unit or operon. A prokaryotic genome has a single site for the initiation of replication, which proceeds bidirectionally; in other words, it is a replicon. Much of our knowledge of the organization of the prokaryotic genome comes from work on the lactose operon of E.coli. Recently, a whole genome, that of the bacteriophage ΦX 174, has been completely sequenced. These investigations have shown that overlapping genes exist in this genome, the same triplets being read in different frames.

The organization of the (nuclear) eukaryotic genome is far more complex. Its replication starts at multiple initiation sites. These multiple sites fulfill an absolute need since, without them, the replication of such a large genome would slow down cell division to an impossibly low rate. The major difference between prokaryotic and eukaryotic genomes, however, is in the organization of genes. Here, the important point is that the generally very large increase in genome size shown by eukaryotes as compared with bacteria is not accom-

panied by a proportionate increase in the number of different proteins encoded. For instance, in the early sea urchin embryo, only 37 million nucleotide pairs appear to be expressed as polysomal messenger RNA; this corresponds to 4% of the haploid genome, or to about 30,000 genes whose average size is equal to 1,200 base pairs. In adult sea urchin tissue, this number drops down to only 2,000 to 5,000 gene equivalents (6,14). This means that the sea urchin DNA expressed in the early embryo is only three times more than that expressed in E.coli, whereas that expressed in adult tissue corresponds to less than half the DNA expressed in E.coli.

These facts raise the problem of the role of the "excess DNA" of eukaryotes. Three general kinds of explanation have been proposed: 1) One possibility is that eukaryotic genes are "serially repeated" up to 1,000 times along the DNA duplex. The nucleotide sequence of each of the repeated or "slave" genes is brought into accord with that of a "master" gene at least once per generation, possibly during the meiotic prophase, by an unspecified rectification mechanism. Because of rectification, only mutational alterations in the master gene are ever detected. This possibility has been disproven as a general model for the organization of the eukaryotic genome on several grounds. It is true, however, that a number of "repeated genes" exist in eukaryotes: ribosomal genes, t-RNA genes, 5S RNA genes, and histone genes. Apart from these, all other genes which have been probed, (globin, ovalbumin, immunoglobulin, fibroin, etc.), have been found to be present in a single copy or a very small number of copies. 2) A second possibility is that the "excess DNA" is largely used for regulatory purposes; as we know, structural genes have not considerably increased in number during evolution, but it is conceivable that the amount of regulatory DNA has increased in amount. If such is the case, one might expect that such regulatory sequences are repeated a large number of times in the genome. 3) Finally, the "excess DNA" might be there for completely different purposes: for main-

tenance of chromosome structure, or as material to be used in evolution, meiosis, recombination, etc. This and the previous possibility are, of course, not mutually exclusive.

The c-value paradox, as the "excess DNA" problem has been called, along with the difficulties of eukaryotic genetics, justifies approaches at the molecular level, in which DNA sequences are directly investigated. Most of the information available so far is derived from the study of the kinetics of DNA renaturation.

THE ORGANIZATION OF THE EUKARYOTIC GENOME AS SEEN BY THE KINETICS OF DNA RENATURATION

The reannealing of separated complementary single strands of DNA ideally follows a second order kinetics. For a given initial DNA concentration, and a certain fragment size, the half-time of reassociation should be proportional to the number of different types of fragments present and thus to the genome size (26,3). This expectation is exactly borne out in the case of the viral and bacterial genomes, which are characterized by a unique copy DNA sequence. Eukaryotic DNAs show complex renaturation kinetics and can usually be resolved into fast, intermediate, and slow-renaturing components (Figure 1). Renaturation kinetics studies, very largely based on a technique developed in our laboratory (13), have contributed two major points to our understanding of the organization of the eukaryotic genome: 1) A large fraction of the genome, usually 50-70%, consists of unique or single-copy sequences. These sequences show a very slow renaturation kinetics. The rest of the DNA renatures at faster rates, since it is made up of repetitive sequences with degrees of reiteration ranging from 10 to 1,000 copies to over 100,000 copies. The latter have been shown to correspond to "satellite" DNAs, which can usually be separated by density gradient centrifugation (see below). Some very fast-renaturing material, following first order

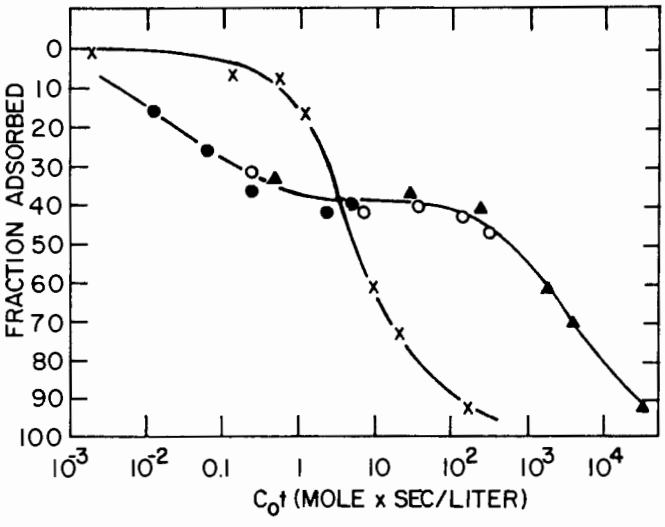

Figure 1. The kinetics of reassociation of calf thymus DNA with hydroxyapatite. Crosses are radioactively labeled E. coli DNA (3).

kinetics, has also been shown to exist in the eukaryotic genome. These fragments can fold back on themselves since they contain palindromic nucleotide sequences; they usually represent a few percent of eukaryotic DNA. 2) The relative arrangement of repetitive and nonrepetitive sequences can be investigated by incubating labeled DNA, sheared to various fragment lengths, with excess 450-nucleotide fragments of unlabeled DNA at a low Cot value (Cot equals the product of initial DNA concentration multiplied by reaction time), allowing reassociation to proceed, and following the binding of labeled DNA to hydroxyapatite (Figure 2). Analysis of data for Xenopus DNA (7) showed that a little more than 50% of the genome consists of closely interspersed repetitive and nonrepetitive sequences. The average length of the repetitive sequence elements is 300 $\pm$ 100 nucleotides, while the nonrepetitive sequences separating adjacent repetitive sequence elements average 800 $\pm$ 200 nucleotides. The remainder of the DNA is mainly nonrepetitive, though most of it contains rare

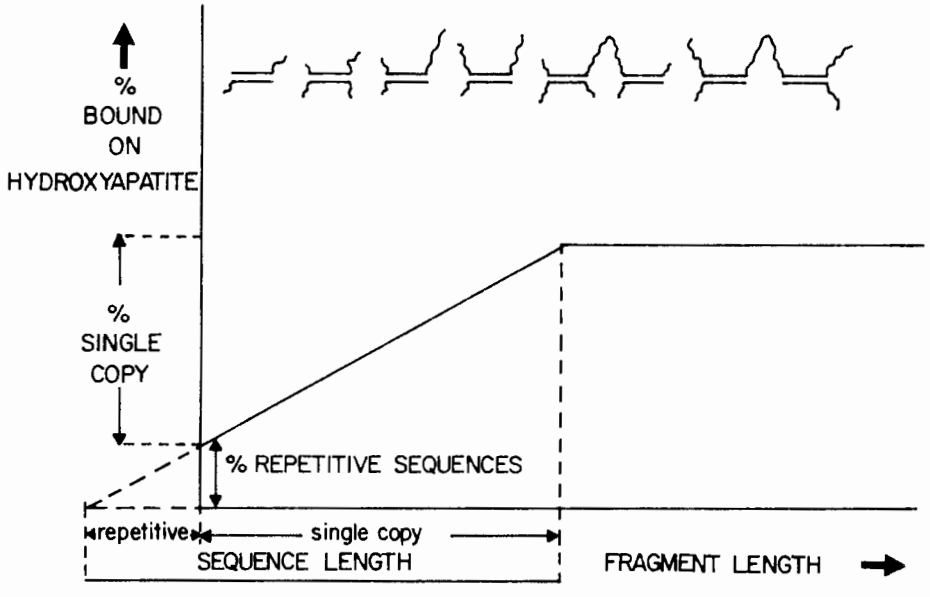

Figure 2. Schematic representation of the effect of fragment size on the fraction of DNA containing repetitive sequence elements, where interspersion is present. The curve results from a model calculation for a uniform length of repetitive sequences interspersed with a uniform length of nonrepetitive sequences (single copy sequences).

Unpaired regions; = paired; paired regions may be terminated by shear or end of repetitive regions (modified from reference 15).

interspersed repetitive elements spaced at a minimum of 4,000 nucleotides apart. A high degree of order in the arrangement of DNA sequences in the Xenopus genome is suggested by the alteration of relatively brief repetitive sequences and longer nonrepetitive sequences; while the latter have the size of single structural genes, the former could serve as sites for the specific binding of regulatory macromolecules. Evidence has been presented that most structural genes expressed in the gastrula of the

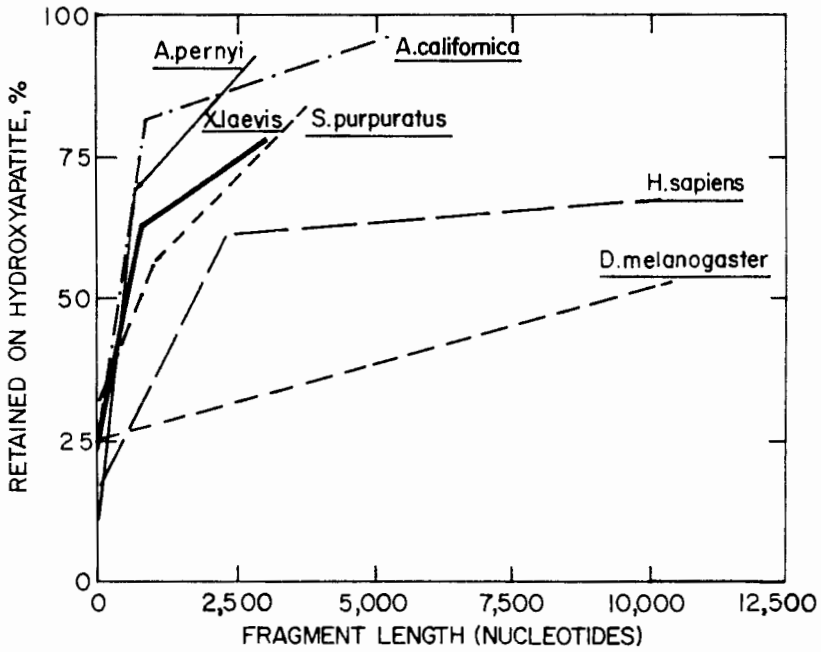

Figure 3. Interspersion data from several eukaryotic genomes (from references 1,7,10,15,19,22).

sea urchin (an organism having the same pattern of genome organization as Xenopus; see Figure 3) are adjacent to repetitive DNA sequences. This is suggested by the ability of single copy sequences adjacent to repetitive sequences to hybridize with labeled gastrular message preparations (8). Interestingly, nuclear RNA from sea urchin gastrulae has a single-copy sequence complexity ten times that expressed in polysomal message and is a transcript of interspersed repetitive and nonrepetitive sequences (23,24). Its function is unknown, but possibly it is of importance to the system of transcription regulation (16).

About 20 species, widely separated phylogenetically and including both deuterostomes and protostomes, show the Xenopus pattern (6,10,14). (Figure 3 and references 1,21). Among insects, one dipteran (Musca domestica) and a lepidopteran (Antherea pernyi) show the Xenopus pattern, while another dipteran (Drosophila melanogaster) and a hymenopteran (Apis mellifera) show a quite different pattern.

There are two major problems with these results. The first is the completely different genome organization exhibited by Drosophila and Apis compared with all other species investigated, which include Musca, a dipteran like Drosophila. The second problem is that the interspersion pattern of eukaryotes as widely separated as protostomes and man appears to be similar, as judged by the kinetics of DNA renatu ration. There is no explanation at present for the first problem. In contrast, the second problem seems to be mainly due to the limitations of the technique, particularly as used on unfractionated eukaryotic genomes.

THE ORGANIZATION OF THE EUKARYOTIC GENOME AS SEEN BY THE ANALYSIS OF DNA IN DENSITY GRADIENTS

It has been well known since the classical work of Meselson, Stahl, and Vinograd (4) that, when centrifuged to equilibrium in CsCl density gradients, bacterial DNAs show unimodal, symmetrical bands, whereas calf thymus DNA exhibits a multimodal, broad, asymmetrical band. Bacterial DNAs, which have essentially statistical nucleotide sequences, show buoyant densities in CsCl density gradients which are linearly related to their base composition (25). This was thought to suggest that the CsCl band profile of calf thymus DNA is due to molecules differing in base composition. It was later realized, however, that the buoyant density in CsCl depends, in fact, not upon base composition, but upon short nucleotide sequences (3,5). Density gradient centrifugation provides, therefore, the basis for a very effective fractionation procedure of complex DNAs like eukaryotic DNAs, particularly when the procedure is performed in the presence of sequence-dependent ligands, like Ag^+, actinomycin, and BAMD [3,6-bis (acetatomercurimethyl) dioxane].

For the reasons just given, density gradient centrifugation methods have led to the separation of 1) "satellite" DNAs, formed by short, repetitive sequences, also called simple-sequence DNAs;

and 2) repetitive genes which contain short, repetitive sequences in the "spacers" separating the transcribed units (ribosomal genes, t-RNA genes, 5S RNA genes, and histone genes).

Both series of DNA components, the simple-sequence DNAs and the repetitive genes, are the best known eukaryotic DNA components. They can be easily prepared and their study by a number of techniques has made big advances in recent years. While the biological role of repetitive genes is well understood, that of simple-sequence DNA components is not. It is known, however, that they are not transcribed and that they are generally localized near the centromeres of chromosomes. It has been suggested that the simple-sequence DNAs play a structural role in meiotic pairing, in chromosome organization, and in recombination.

Both "satellite" DNAs and repetitive genes generally represent just a few percent of total eukaryotic DNA. Therefore, it is evident that all the unsolved problems reside in the so-called "main-band" DNA. Using centrifugation of DNA in Cs_2SO_4-Ag^+ density gradients, we not only found that we could separate four "satellite" components and several minor components, but we also discovered four years ago (9) that the main band of calf DNA, (molecular weight 13 million), was not formed by genome fragments having a continuous distribution of their average G+C-content, as was commonly believed, but by three distinct families of fragments. These were characterized by buoyant densities in CsCl equal to 1.697, 1.704, and 1.709 g/cm^3, and represented about 50, 25, and 10% of the DNA respectively (Figure 4). These three components gave rise to symmetrical CsCl bands, which exhibited standard deviations close to those of bacterial DNAs of comparable molecular weight; these components were mainly responsible for the asymmetry of the CsCl main band of calf DNA.

Three important questions were raised by the presence of these major DNA components in the calf genome: their evolutionary origin, their relative arrangement and their intramolecular heterogeneity. A detailed analysis of DNA preparations from 25

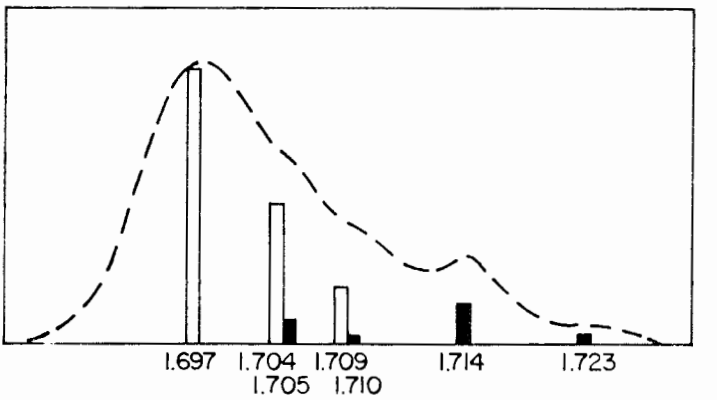

Figure 4. A histogram of the DNA components of the bovine genome. The height of the bars is proportional to the percentage of each component, filled bars correspond to the sharp melting; empty bars to the broad melting components. The broken line is an enlarged band pattern in CsCl of total calf DNA (1).

eukaryotes (23), ranging from yeast to man, led us to the recognition of phylogenetic differences at the macromolecular level in the organization of eukaryotic genomes. The differences concerning the main band DNA are: 1) the three major components, first observed in calf DNA, were also found in the other ten mammalian DNAs investigated; similar components appear to be present in avian genomes; 2) the DNAs of reptiles, amphibians, and fish showed a much lower and decreasing asymmetry of their CsCl bands compared to both mammalian and avian DNAs; 3) essentially symmetrical bands in CsCl were exhibited by three invertebrate DNAs; 4) the DNAs from three unicellular eukaryotes exhibited perfectly symmetrical bands in CsCl and could not be resolved into discrete components (Figures 5,6).

A novel feature of the main component of some mammalian DNAs is that it is made up of two subcomponents, which significantly differ in both buoyant densities and relative amounts in species which are

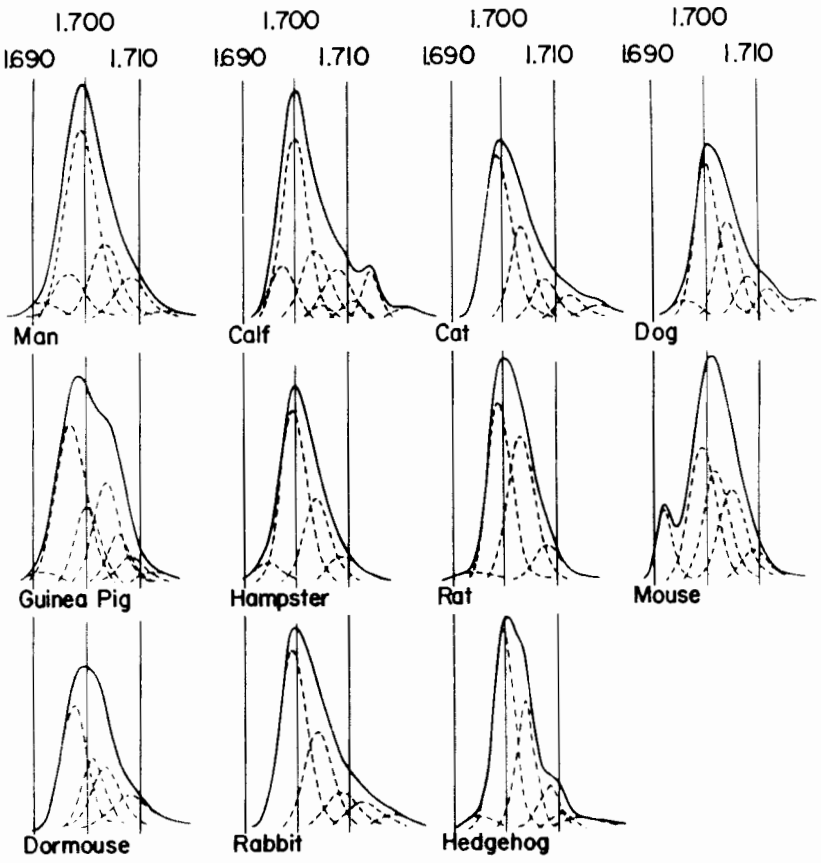

Figure 5. CsCl analysis of DNAs from 11 mammals (24).

evolutionarily not far from each other, and, more strikingly, in different tissues from the same species. Such differences are not due to differences in breed or sex since they have been seen in tissues from a single calf fetus. Two important open problems concerning the 1.704 and 1.709 g/cm^3 components seen in mammals and birds are: first, whether they are homologous in the two orders; and, second, what their biological significance is, e.g., whether they are associated with new functions like homeothermy or amnios formation.

The other two questions were answered by studying (19) a mammalian DNA in the molecular weight range of 1 x 10^6 to more than 200 x 10^6.

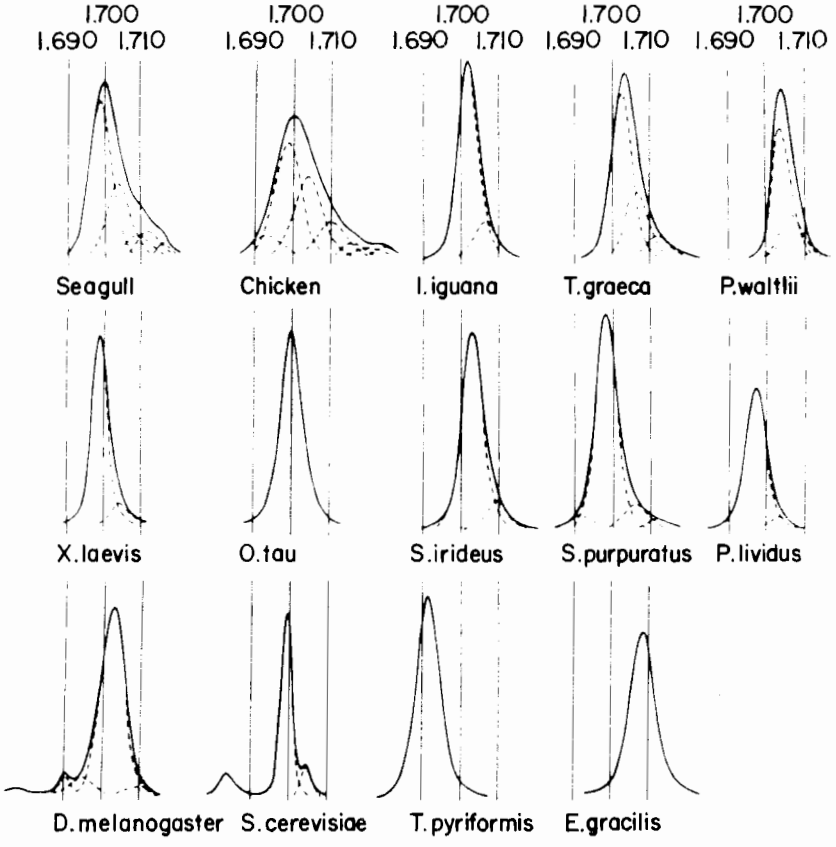

Figure 6. CsCl analysis of DNAs from birds, reptiles, amphibia, fishes, invertebrates, and unicellular eukaryotes (24).

The relative amounts and the buoyant densities of the major components did not vary in this size range indicating that the components are not intermingled with each other over very great lengths of DNA and that their internal sequence heterogeneity is not yet apparent at rather low molecular weight. In contrast, the major components of degraded amphibian DNAs showed a very different pattern compared with the undegraded DNAs, particularly in the case of amphibia having a very large genome size; changes in buoyant densities and amounts of some components, the disappearance of some components,

and the appearance of new components were observed. Finally, the degraded Drosophila and Saccharomyces DNAs were similar to the undegraded preparations in that no components could be resolved within the main bands.

In conclusion, density gradient fractionation of native eukaryotic DNAs is an extremely powerful technique which probably provides the best first step to study the organization of eukaryotic genomes. Its possibilities are far from being fully explored, as shown by our recent work on the calf genome (2,20) combining Cs_2SO_4-BAMD and Cs_2SO_4-Ag^+ density gradient centrifugation, which has led to the preparation of seven simple-sequence and ten minor components. Obviously, this approach can be combined with a number of other techniques like renaturation kinetics studies, restriction enzyme degradation, cloning of DNA fragments, and hybridization experiments.

REFERENCES

1. Angerer, R.C., Davidson, E.H., and Britten, R.J.:
Cell, 6: 29, 1975.

2. Bernardi, G.:
Nature, 206: 779, 1965.

3. Britten, R.J., and Kohne, D.E.:
Science, 161: 529, 1968.

4. Corneo, G., Ginelli, E., Soave, C., and Bernardi, G.:
Biochemistry, 7: 4373, 1968.

5. Cortadas, J., Macaya, G., and Bernardi, G.:
Eur J Biochem (in press).

6. Crain, W.E., Davidson, E.H., and Britten, R.J.:
Chromosoma, 59: 1, 1976.

7. Davidson, E.H., Hough, B.R., Amerson, C.S., and Britten, R.J.:
J Mol Biol, 77: 1, 1973.

8. Davidson, E.H., Hough, B.R., Klein, W.H., and Britten, R.J.:
Cell, 4: 217, 1975.

9. Davidson, E.H., Klein, W.H., and Britten, R.J.:
Dev Biol, 54 (in press).

10. Efstratiadis, A., Crain, W.R., Britten, R.J., Davidson, E.H., and Kafatos, F.C.:
Proc Nat Acad Sci USA, 73: 2289, 1976.

11. Filipski, J., Thiery, J.P., and Bernardi, G.:
J Mol Biol, 80: 177, 1973.

12. Galau, G.A., Britten, R.J., and Davidson, E.H.:
Cell, 2: 9, 1974.

13. Galau, G.A., Klein, W.H., Davis, M.M., Wold, B.J., Britten, R.J., and Davidson, E.H.: Cell, 7: 487, 1976.

14. Goldberg, R.B., Crain, W.R., Ruderman, J.V., Moore, G.P., Barnett, T.R., Higgins, R.C., Gelfand, R.A., Galau, G.A., Britten, R.J., and Davidson, E.H.: Chromosoma, 51: 225, 1975.

15. Graham, D.E., Neufeld, B.R., Davidson, E.H., and Britten, R.J.: Cell, 1: 127, 1974.

16. Hough, B.R., Smith, M.J., Britten, R.J., and Davidson, E.H.: Cell, 5: 291, 1975.

17. Macaya, G., Thiery, J.P., and Bernardi, G.: J Mol Biol, 108: 237, 1976.

18. Macaya, G., Cortadas, J., and Bernardi, G.: Eur J Biochem (submitted).

19. Manning, J.E., Schmid, C.W., and Davidson, N.: Cell, 4: 141, 1975.

20. Meselson, M., Stahl, F.W., and Vinograd, J.: Proc Nat Acad Sci USA, 43: 582, 1957.

21. Schildkraut, C.L., Marmur, J., and Doty, P.: J Mol Biol, 4: 430, 1962.

22. Schmid, C.W., and Deininger, P.L.: Cell, 6: 345, 1975.

23. Smith, M.J., Hough, B.R., Chamberlin, M.E., Britten, R.J., and Davidson, E.H.: J Mol Biol, 85: 103, 1974.

24. Thiery, J.P., Macaya, G., and Bernardi, G.: J Mol Biol, 108: 219, 1976.

25. Wells, R.D., and Blair, J.E.:
J Mol Biol, 27: 273, 1967.

26. Wetmur, J.G., and Davidson, N.:
J Mol Biol, 31: 349, 1968.