

## SEQUENCE ORGANIZATION OF THE VERTEBRATE GENOME

Giorgio Bernardi

Laboratoire de Génétique Moléculaire, Institut Jacques  
Monod, 2 Place Jussieu, 75005 Paris, France.

## ABSTRACT

The sequence organization of the nuclear genome of vertebrates has been studied by density gradient centrifugation in the presence of DNA ligands. This approach has been combined with renaturation kinetics and with studies of specific sequences using cloned probes. A new picture of the vertebrate genome has emerged from these investigations.

---

A number of impacts of genetic manipulations on man and society have been discussed in this Symposium. I would like to suggest that yet another one may turn out to be most important. This is our understanding of the organization, of the informational content, and of the functional regulation of the human genome. At the present time, we are still far from this goal, simply because of the formidable complexity of the problems involved. Yet, the technology needed to solve them is now available and one can foresee that we will reach, perhaps in a relatively near future, this ultimate level of the knowledge of ourselves, understanding what determines to such a large extent what we are, namely our own genome. When we will reach this goal, we will fulfill the ancient Greek precept "know yourself".

Here I would like to summarize very briefly investigations carried out in my laboratory on this subject over the past fifteen years. Our studies have not been limited to the human genome, because we thought that a broader coverage of vertebrate genomes could provide us with useful evolutionary insights.

The basic experimental approach we have used, equilibrium centrifugation of native DNA in density gradients containing DNA ligands, has a very solid theoretical background and can be combined with both DNA reassociation studies and recombinant DNA technology. The resolving power of this approach depends upon differences in the frequencies of short (oligonucleotidic) sequences in the DNA fragments under study. This sequence-dependence, already evident in CsCl (Wells and Blair, 1967; Corneo *et al.*, 1968), is enhanced in the presence of a number of DNA ligands and leads, for instance, to a high resolution of satellite and ribosomal DNAs (Corneo *et al.*, 1968; Filipski *et al.*, 1973; Cortadas *et al.*, 1977; Macaya *et al.*, 1978; Meunier-Rotival *et al.*, 1979). More interestingly, however, we discovered that density gradient centrifugation allows the reso-

lution of mammalian and avian main-band DNAs into a small number of discrete components, which could, therefore, be studied in detail, so providing a wealth of new information.

Neglecting satellite and minor components, mammalian and avian genomes can be resolved (Filipski et al., 1973; Thierry et al., 1976; Macaya et al., 1976; Cortadas et al., 1979) in  $\text{Cs}_2\text{SO}_4/\text{Ag}^+$  or  $\text{Cs}_2\text{SO}_4/\text{BAMD}$  (BAMD is (acetato-mercuri-methyl)dioxane), into: a) one or two (according to the species) light components, representing about 2/3 of the genome, and having buoyant densities in the 1.697-1.703  $\text{g}/\text{cm}^3$  range; and b) two heavy components, representing about 25% and 10% of the genome, and having buoyant densities of 1.704 and 1.708  $\text{g}/\text{cm}^3$ , respectively. In contrast, cold-blooded vertebrates have main-band DNAs ranging in density, in most species, from 1.697 to 1.703  $\text{g}/\text{cm}^3$  (which is the range of the light components of warm blooded DNAs) and characterized (like those of unicellular organisms and invertebrates) by CsCl band profiles which are symmetrical or only slightly skewed on the heavy side (Thierry et al., 1976; Macaya et al., 1976); expectedly, DNAs from warm-blooded vertebrates exhibit CsCl bands strongly skewed on the heavy side. The symmetry properties of CsCl bands of vertebrate DNAs are paralleled by their compositional heterogeneities, as assessed by the width of CsCl bands; for instance, the compositional heterogeneities of a number of main-band DNAs of fish (Hudson et al., 1980) are as small or smaller than those of individual major components from warm-blooded vertebrates (Cuny et al., 1981; Olofsson and Bernardi, 1983).

The finding of a small number of discrete major components in the genomes of warm-blooded vertebrates raises two questions, namely a) whether the heavy components correspond to newly formed sequences which did not exist in cold-blooded vertebrates, or to sequences which did exist in cold-blooded vertebrates, but underwent an increase in GC; and b) the molecular weight range in which the DNA fragments forming the major components can be separated.

The first question is clearly answered by the finding that a number of genes localized in the main bands of cold-blooded vertebrates (which correspond in buoyant density to the light components of warm-blooded vertebrates), are present in the heavy components of warm-blooded vertebrates. Such is the case, for instance, of the mouse  $\alpha$ -globin gene and of the chicken  $\alpha$ - and  $\beta$ -globin genes, which are all localized on 1.708 components (Bernardi, 1979; and paper in preparation), whereas in *Xenopus* they are localized on main-band DNA fragments having a buoyant density of 1.700  $\text{g}/\text{cm}^3$  (G. Cuny, pers. comm.). In other words, the heavy components of warm-blooded vertebrate DNAs are the result of regional increases in GC, (which affect both coding and non-coding DNA sequences, at it will be shown below), and not the result of de novo formation of sequences.

As far as the second question is concerned, since the major components do not vary in their relative amounts nor in their buoyant

densities, whether they are obtained from DNA preparations as low as  $2.10^6$  or higher than  $2.10^8$  in molecular weight, (Macaya *et al.*, 1976), one has to conclude that they are formed by families of fragments derived from very long chromosomal segments endowed with a remarkable compositional homogeneity, which we have called isochores. This conclusion is supported by the fact that all single-copy or clustered genes tested so far (about 20) have always been found on one single major component (a proof, by itself, of a real fractionation), even when the molecular weight of the fractionated DNA preparation was in excess of  $50.10^6$  (or 75 kb). Since these genes could be localized anywhere on the fragments derived by random breakage, the regions of compositional homogeneity must be in excess of  $100.10^6$  (or 150 kb).

Several points suggest (Cuny *et al.*, 1981) that isochores may correspond to G- and R-bands of chromosomes : a) there are strong indications that G-bands correspond to AT-rich, late-replicating DNA, and R-bands to GC-rich, early replicating DNA; chromosome bands would therefore differ in GC contents like isochores; b) G-bands are very evident in warm-blooded, but faint or absent in cold-blooded vertebrates, a feature paralleled by the very different compositional heterogeneities of DNAs from warm-blooded and cold-blooded vertebrates; c) no bands can be detected in chromosomal regions harboring an amplified gene, as expected from the fact that the amplified region is small compared to an isochore; d) the amount of DNA per chromosomal band, even taking into account high resolution banding, is more than compatible with the estimated isochore size; e) G-bands and R-bands are interspersed in chromosomes and so are isochores, as shown by the fact that genes belonging to scattered multigene families, like the actin genes, are found on all major components of mouse and human DNAs (Soriano *et al.* 1982) and also at 30 different loci on metaphase chromosomes; f) G-banding patterns appear to be highly conserved in warm-blooded vertebrates, as are the relative amounts and the buoyant densities of major components as well as the localization of genes within individual major components; for example, the  $\beta$ -globin genes of mouse, man and rabbit are all on one light component, whereas the  $\alpha$ -globin genes of the same species are on the heaviest component.

Concerning the problem of sequence organization, the reassociation kinetics of DNA fragments derived from isolated major components from mouse, man and chicken (Soriano *et al.*, 1981; Olofsson and Bernardi, 1983) has shown that the relative amounts of interspersed repeated and unique sequences (as well as foldback sequences in the case of the mammalian genomes) strikingly differ in the different major components of each genome; for example, most of the interspersed sequences of chicken are concentrated in the 1.708 component, which is the poorest in such sequences in the mouse genome. These findings lead to the general conclusion that the sequence organization of mammalian and avian genomes is not uniform in different chromosomal regions and, furthermore, that it exhibits remarkable differences in different mammals and between mammals and birds. Contrary to the cliché of a universal Xenopus pattern, a wide

variety of sequence arrangements exist already within vertebrates and within mammals. Furthermore, because of the correlation between major components and isochores, one should consider beside the short range (1-10 kb) distribution of repeated sequences (corresponding to the classical interspersion patterns), a long-range (> 100 kb) distribution of repeated sequences due to the interspersion of isochores and to their different internal distribution of repeated sequences.

A closer look at the distribution of interspersed repeats can be obtained by using cloned repeated sequences as probes. Three families of short repeats (Alu, B1 and CRI; Schmid and Jelinek, 1982; Stumph *et al.*, 1982) and two families of long repeats (Bam and Kpn; Meunier-Rotival *et al.*, 1982; Soriano *et al.*, 1983) were investigated and found to be concentrated in the heavy components (particularly the 1.708 component) and in the light components, respectively of the DNAs of man, mouse and chicken. These results are of interest because : a) they show a very similar if not identical distribution for specific families of repeats, which are known to share sequence homology; b) in no case the distribution of these families is ubiquitous; at least one major component and often two components "contiguous" in density are either totally free or very low in some repeated sequences. As a consequence, the postulated mobility of the repeats must be accompanied by strong constraints on the use of integration sites, which can only be located in some major components but not in others. Likewise, the integration of the bovine leukemic proviral DNA was shown to occur at a number of sites of the host cell genome, all of which are, however, localized in the 1.708 component (Kettmann *et al.*, 1979).

A final point concerns the correlation between the base composition of repeated sequences and that of the major components in which they are embedded. In the cases studied so far, the correlation is excellent, the short repeats present in the heavy components being higher in GC than the long repeats present in the light components; in both cases, the match between the compositions of repeats and corresponding components is remarkable (Soriano *et al.*, 1983). A similar correlation is found with genes and even with gene segments (untranslated and translated regions, introns) some of which, (the introns), are known to undergo changes by insertion and deletion rather than by point mutation. These results agree with the finding that, within each major component, interspersed repeated sequences and unique sequences do not differ in their GC contents (Soriano *et al.*, 1981). Needless to say that the match between coding sequences and major components in base composition can only be obtained if constraints on codon usage exist; in fact, these have been detected (Bernardi, 1979; and paper in preparation). Genes located in heavy components in contrast with those located in light components show a preference for codons rich in G and C. Interestingly, this leads to high levels of the doublet CG, usually avoided in vertebrate genomes, in those genes.



In conclusion, the discovery of isochores and the study of the properties of DNA fragments derived from them puts the problem of the sequence organization of vertebrate genome in a new perspective, revealing features unsuspected so far, like the compartmentalization of sequences into chromosomal domains, and the different sequence organizations of genomes belonging to species relatively close to each other from an evolutionary viewpoint. Investigations on isochores should help in bridging the gap between studies at the gene and at the chromosomal level. Finally, the question of the origin of isochores is raised; it is possible, that the isochores of warm-blooded vertebrates have arisen for purely structural reasons; if such is the case, these reasons must be really compelling, since they appear to go as far as to impose constraints on codon usage.

## REFERENCES

- Bernardi, G. (1979). Organization and evolution of the eukaryotic genome. in *Recombinant DNA and Genetic Experimentation* (Morgan J. and Whelan W.J., eds.) pp. 15-19, Pergamon.
- Corneo, G., Ginelli, E., Soave, C., and Bernardi, G. (1968). Isolation and characterization of mouse and guinea pig satellite DNA's. *Biochemistry* 7, 4373-4379.
- Cortadas, J., Macaya, G., and Bernardi, G. (1977). An analysis of the bovine genome by density gradient centrifugation: fractionation in  $Cs_2SO_4$ / 3,6 bis (acetato-mercuri-methyl) dioxane density gradient. *Eur. J. Biochem.* 76, 13-19.
- Cortadas, J., Olofsson, B., Meunier-Rotival, M., Macaya, G., and Bernardi, G. (1979). The DNA components of the chicken genome. *Eur. J. Biochem.* 99, 179-186.
- Cuny, G., Soriano, P., Macaya, G., and Bernardi, G. (1981). The major components of the mouse and human genomes: preparation, basic properties, and compositional heterogeneity. *Eur. J. Biochem.* 115, 227-233.
- Filipski, J., Thiery, J.P., and Bernardi, G. (1973). An analysis of the bovine genome by  $Cs_2SO_4$ - $Ag^+$  density gradient centrifugation. *J. Mol. Biol.* 80, 177-197.
- Hudson, A.P., Cuny, G., Cortadas, J., Haschemeyer, A.E.V. and Bernardi, G. (1980). An analysis of fish genomes by density gradient centrifugation. *Eur. J. Biochem.* 112, 203-210.
- Kettmann, R., Meunier-Rotival, M., Cortadas, J., Cuny, G., Ghysdaee, J., Mammericks, M., Burny, A., and Bernardi, G. (1979). Integration of bovine leukemia virus DNA in the bovine genome. *Proc. Natl. Acad. Sci. U.S.A.* 76, 4822-4826.

- Macaya, G., Cortadas, J., and Bernardi, G. (1978). An analysis of the bovine genome by density-gradient centrifugation. *Eur. J. Biochem.* 84, 179-188.
- Macaya, G., Thiery, J.P., and Bernardi, G. (1976). An approach to the organization of eukaryotic genomes at a macromolecular level. *J. Mol. Biol.* 108, 237-254.
- Meunier-Rotival, M., Cortadas, J., Macaya, G., Bernardi, G. (1979). Isolation and organization of calf ribosomal DNA. *Nucleic Acids Res.* 6, 2109-2123.
- Meunier-Rotival, M., Soriano, P., Cuny, G., Strauss, F., and Bernardi, G. (1982). Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. *Proc. Natl. Acad. Sci. U.S.A.* 79, 355-359.
- Schmid, C.W., and Jelinek, W.R. (1982). The Alu family of dispersed repetitive sequences. *Science* 216, 1065-1070.
- Soriano, P., Macaya, G., and Bernardi, G. (1981). The major components of the mouse and human genomes : reassociation kinetics. *Eur. J. Biochem.* 115, 235-239.
- Soriano, P., Meunier-Rotival, M., and Bernardi, G. (1983). The distribution of interspersed repeats is non-uniform and conserved in the mouse and human genomes. *Proc. Natl. Acad. Sci. U.S.A.* 80, 1816-1820.
- Soriano, P., Szabo, P. and Bernardi, G. (1982). The scattered distribution of actin genes in the mouse and human genomes. *EMBO J.* 1, 579-583.
- Stumph, W.E., Kristo, P., Tsai, M.-J., and O'Malley, B.W. (1981). A chicken middle repetitive DNA sequence which shares homology with mammalian ubiquitous repeats. *Nucleic acids Res.* 9, 5383-5397.
- Thiery, J.P., Macaya, G., and Bernardi, G. (1976). An analysis of eukaryotic genomes by density gradient centrifugation. *J. Mol. Biol.* 108, 219-235.
- Wells, R.D. and Blair, J.E. (1967). Studies on polynucleotides. LXXI. Sedimentation and buoyant density studies of some DNA-like polymers with repeating nucleotide sequences. *J. Mol. Biol.* 27, 273-288.

## DISCUSSION

A..E. SIPPEL: Did you measure C5-methylation percentages in your four components? I ask this because you mentioned different degrees of CG elimination in your components.

G. BERNARDI: This was tested by comparing Hpa II and Msp I digests of isolated components, but this method is not sensitive enough to detect the expected changes.

N. ZINDER: Can you relate base-compositional differences to chromosomes?

G. BERNARDI: All components seem to be represented on all chromosomes.