ORGANIZATION AND EVOLUTION OF THE EUKARYOTIC GENOME

G. Bernardi

**Laboratoire de Génétique Moléculaire,
Institut de Recherche en Biologie Moléculaire,
Université Paris VII, 2 Place Jussieu, 75005 Paris, France**

The organization of the prokaryotic genome and the regulation of its expression are reasonably well understood at the present time. In contrast, these problems are still quite open in the case of the eukaryotic genome, in spite of the efforts of many laboratories in this area during the past few years. This situation is that much more regrettable since the issues under consideration are of capital importance for understanding evolution and differentiation.

In the present brief review I will attempt, first, to introduce the major questions concerning the organization of the eukaryotic genomes (without touching the problem of its regulation); second, to discuss what we have learned from the experimental approach which has been most widely used in recent years, namely the kinetics of DNA renaturation ; and third, to present a different approach to the problem.

Living organisms present a major discontinuity, separating prokaryotes and eukaryotes ; no intermediate forms are known. A comparison of prokaryotes and eukaryotes reveals that the major differences concern the size, the structure and the organization of the genome.

The genome size, namely, the amount of (nuclear) DNA per haploid cell, is constant in each eukaryotic species and covers an extremely wide range of values. Some fungi have genome sizes practically equal to those of some bacteria (the latter have a very limited range of genome sizes), 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 size are often found within single orders, within single genera and even within interbreeding species. Since it is unlikely that these differences correspond to comparable differences in the amount of genetic information, the minimum genome sizes found in each order are usually considered, neglecting the interesting but less important problem of the variation of genome size within orders. Even so, a ratio of about 1000 is found between the smallest genome size of prokaryotes and the largest (minimum) genome size found in eukaryotes, that is, the genome size of mammals. A well-defined trend exists for the minimum genome size to increase with evolution. Very interestingly, if one plots such minimum genome size against the divergence time of different orders, it becomes evident that one can distinguish two phases in organ evolution: one

in which genome size varied very little and one in which genome size strikingly increased. Very interestingly, the separation between these two phases corresponds to the appearance of multicellular organisms and of cellular differentiation.

The genome structure of eukaryotes is much more complex than that of prokaryotes. We will not discuss this point here. Suffice it to mention that at least three distinct structural levels have been recognized in eukaryotes, that of nucleosomes, that of chromomeres, and that of chromosomes. Another major difference between prokaryotic and eukaryotic cells is the segregation, in eukaryotes, of part of the genome into cytoplasmic organelles (mitochondria, chloroplasts).

Concerning the organization of the eukaryotic genome, the fundamental point here is that the genome size increase occurring in evolution has not been accompanied by a corresponding increase in the number of different polypeptide chains encoded. In general, it can be said that only a small percentage of the eukaryotic genome is expressed. For instance, in the early sea urchin embryo, only 4 % of the haploid genome appears to be expressed as polysomal mRNA ; in adult sea urchin tissues, this number drops to less than 1 % of the genome, a value lower than that of the DNA expressed in E.coli. These data stress what certainly is the most striking difference existing between the prokaryotic and the eukaryotic genomes. The former is made up simply of genes transcribed into mRNAs, rRNAs and tRNAs and of short regulatory sequences preceding each polycistronic transcription unit.The latter contains a large excess of DNA, compared to what is found in the final transcripts. Only a fraction of this excess DNA is accounted for. First of all,a certain percentage of eukaryotic DNA is present in simple highly repeated sequences, forming what are known as satellite DNAs ; these DNA segments are not transcribed and have a function which is still unknown. Second, some genes e.g. the rRNA, tRNA, and histone genes are present in multiple copies. Third, a number of eukaryotic genes contain non-coding sequences, the so-called intervening sequences, which may represent as much as 5 to 10 times the amount of DNA contained in the corresponding coding sequences.

The majority of the excess DNA is, however, not accounted for yet. The fact that most of it is in all likelihood non-coding has encouraged approaches in which the genome organization of eukaryotes is studied directly at the molecular level.

THE KINETICS OF RENATURATION OF EUKARYOTIC DNA

The main experimental approach used so far has been the study of the kinetics of renaturation exhibited by DNA fragments. The reannealing of separated complementary single strands of DNA ideally follows second order kinetics. For a given initial DNA concentration and a certain DNA 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. This expectation is exactly borne out in the case of viral and bacterial genomes, which are characterized by a unique DNA sequence. Eukaryotic DNAs, in contrast, show complex renaturation kinetics and can usually be resolved into fast, intermediate and slow-renaturing components. The latter represent in most cases 50-70 % of the genome, are formed by single copy sequences and comprise most eukaryotic genes and their intervening sequences ; the intermediate DNA is made up of repetitive sequences with degrees of reiteration comprising between 10 and 1000 copies, the fast DNA corresponds to satellite DNAs, the sequences of which are repeated over 100,000 times. In addition, some very fast renaturing material, following first order 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.

The relative arrangement of repetitive and non-repetitive (single-copy) sequences was investigated by reassociating to a low cot,(the product of the initial DNA concentration by renaturation time), labelled DNA, sheared to various fragment lengths, with excess short fragments of unlabelled DNA and by following the binding of labelled DNA to hydroxyapatite. Such analysis as applied to the Xenopus genome has shown that about 50 % of this DNA consists of closely interspersed repetitive and non-repetitive sequences (short-period interspersion). The average length of the repetitive sequence elements is 300 + 100 nucleotides, while the non-repetitive sequences separating adjacent repetitive sequence elements average 800 + 200 nucleotides. The remainder of DNA is mainly non-repetitive, though most of it contains rare interspersed repetitive elements spaced at a minimum of 4000 nucleotides apart (long-period interspersion).

Over 20 species, widely separated phylogenetically, have been shown to be endowed with the Xenopus interspersion pattern ; among insects, one dipteran(Musca domestica)and one lepidopteran (Antherea pernyi) show the Xenopus pattern, while another dipteran,(Drosophila melanogaster) and a hymenopteran (Apis mellifera) show a quite different pattern in which the repeated sequences are much more widely spaced from each other than in the short-period interspersion of Xenopus.

In summary, it can be said that the major contribution of this approach to our understanding of the organization of the eukaryotic genomes has been the demonstration that these genomes contain, in contrast to prokaryotic genomes,repeated sequences which are interspersed with the unique sequences. It has been speculated that the characteristic interspersed repeated sequences are correlated with the regulation of the expression of eukaryotic genes. Such speculation does not have, however, any experimental support and seems to be contradicted by the interspersion patterns of insect genomes. An alternative hypothesis, which seems more reasonable, is that the interspersed repeated sequences play a role in the unequal crossing-over phenomena, which were responsible, in all likelihood, for the process of evolutionary increase in genome size exhibited by eukaryotes. In any event, it seems that renaturation kinetics, at least as applied to unfractionated eukaryotic genomes, has provided all the information it can give and that new approaches are needed.

DENSITY GRADIENT FRACTIONATION OF EUKARYOTIC DNA

The approach which has been mainly used in our laboratory is based on the fractionation of eukaryotic DNA by density gradient centrifugation, in the presence of DNA ligands, mainly Ag^+ and an organic mercurial, bis-(acetato-methylmercuri) dioxane or BAMD. These techniques separate, in general, native DNA fragments containing short repeated nucleotide sequences according to their sequences and other DNA fragments according to base composition. Satellite DNAs and repeated genes, which have satellite-like sequences in their spacers,are easily separated, in general. For this reason, from now on, we will disregard them and consider the fractionation of the bulk of eukaryotic DNA, the so-called main-band DNA.When studying DNAs from eukaryotes widely distant from a phylogenetic point of view, we observed that symmetrical CsCl bands were exhibited by unicellular eukaryotes and invertebrate DNAs, as is the case for prokaryotic DNAs; the DNAs from fishes, amphibia and reptiles exhibited a very slight and increasing asymmetry on the heavy side of their CsCl bands ; the DNAs from warm-blooded vertebrates, birds and mammals, exhibited CsCl bands which were very asymmetrical on the heavy side. A fine analysis involving density gradient centrifugation in the presence of Ag^+ or BAMD led us to the recognition of 4 discrete DNA components in the main band of avian and mammalian DNAs. The existence of these discrete components has been confirmed by preparing them. The major DNA components exhibit, when run in CsCl, gaussian bands and a compositional heterogeneity very close to that of bacterial

DNAs. The relative amount and the buoyant densities of the major components of mammalian and avian genomes are very close to each other. In the case of the mouse genome, the four major components have buoyant densities equal to 1.699, 1.701, 1.704 and 1.708 g/cm^3 and represent about 26 %, 35 %, 18 % and 8 % of the genome, respectively. It should be noted that the major light DNA components of avian and mammalian DNAs are in the same buoyant density range as the DNAs of cold-blooded vertebrates, and that the major heavy components are responsible for the asymmetry of their CsCl bands.

Again in the case of the mouse genome, the renaturation kinetic properties show that most of the fold-back and interspersed repetitive sequences are present in the two light components, the two heavy ones being mainly formed by single copy sequences. In all cases investigated, the major components of warm-blooded vertebrates represent blocks of rather homogeneous base composition which are larger in molecular weight than 100 million daltons.

The existence of discrete major components in the genomes of warm-blooded vertebrates is of interest because it implies that different sections of these genomes are under different compositional constraints. While the reasons for such a situation are not yet clear, it may be relevant in this connection to mention very recent results of G. Cuny, M. Meunier and P. Soriano of our laboratory on the location of globin genes in the DNA components of rabbit, mouse and man (probes obtained from T. Maniatis, C. Weissmann, and B. Williamson were used). In all cases, the β-globin gene was found to be present in the 1.701 component ; preliminary results indicate that this is also the location of the human γ-globin gene, which is contiguous to the β -globin gene. In contrast, the α-globin gene has been localized in the 1.708 component of the mouse genome. These results are interesting for two main reasons : 1) α-globin and β-globin genes are the result of a gene duplication; a translocation of one of the two genes took place at a certain point in evolution, as witnessed by the different chromosomal and component location of the two genes. Now, the component location, but not the chromosomal location, indicates that it was the α -globin gene which was translocated from its original position; in fact, not only does the 1.708 component not exist in lower vertebrates, but also no DNA fragment having such a high density is detected. In contrast, the further duplications of the human β-globin gene remained in the component where the β-globin gene was and still is located, as witnessed by the location of the γ-globin gene. 2) The base composition of α- and β -globin mRNAs from rabbit, mouse and man are known, as well as that of human γ-globin genes. The α -globin mRNAs have a remarkably higher GC content (64 % for man and rabbit) than the β -globin mRNAs (51% for man, rabbit and mouse) and the γ-globin mRNA (51% for man). This is a surprising result if one considers that α-globin mRNA could have the same base composition as β-globin mRNA, and that its GC content has been increased at the price of having a large number of otherwise forbidden or avoided GC doublets. Considering that all intervening sequences studied so far have a lower GC content than the coding sequences (the β -globin gene of mouse has, for instance, 46 % GC versus 51 % GC for its mRNA), it is likely that the GC contents of globin genes are close to the average base composition of the large DNA blocks in which they are embedded. If this conclusion is confirmed and extended to other genes, the compositional constraints seen to exist in the major components of the eukaryotic genome extend to the genes they contain.

REFERENCES

Concerning the kinetics of DNA renaturation , the reader is referred to the papers published by Britten, Davidson, and their colleagues. The analysis of globin cDNA has been reported by Konkel et al. (Cell, 15. 1125-1132, 1978) and

by Forget et al.at the ICN-UCLA Symposium on Eukaryotic Gene Expression (March 1979 ; paper in press). Previous papers from our own laboratory are given below.

Corneo G., Ginelli E., Soave C. and Bernardi G.
Biochemistry, 7. 4373 (1968)

Filispki J., Thiery J.P. and Bernardi G.
J. Mol. Biol., 80. 177 (1973)

Thiery J.P., Macaya G. and Bernardi G.
J. Mol. Biol., 108. 219 (1976)

Macaya G., Thiery J.P. and Bernardi G.
J. Mol. Biol.108. 237 (1976)

Cortadas J., Macaya G. and Bernardi G.
Eur. J. Biochem., 76. 13-19 (1977)

Macaya G., Cortadas J. and Bernardi G.
Eur. J. Biochem., 84. 179-188 (1978)

Cuny G., Macaya G., Meunier-Rotival M., Soriano P. and Bernardi G.
in Genetic Engineering (H.W. Boyer and S. Nicosia, eds.)
Elsevier, Amsterdam, 1978, pp. 109-115.

DISCUSSION

J.D. WATSON: On the point about intervening sequences, I don't think it's widely known that you can get great variations in DNA content ("C" values) not only over a long evolutionary period, but even within plant genera where you can get factors of ten variation. This was first emphasised by Stebbens and re-discovered by Joshua Lederberg, who was intrigued by the fact that plants that have a relatively small amount of DNA have the same number of chromosomes as those with tenfold more DNA. So, it's not a question of polyploidy lost in patches. Those species which have very short life cycles have the small DNA content, whereas those which have lots of DNA have long life cycles. If one combines these facts with the observation from Tonegawa's laboratory - that the intervening sequences occur between functional domains - the speculation arises as to whether the intervening sequences largely serve to promote recombination between functional domains. Perhaps the main reason for the vast increase in DNA is to promote recombination, as evolution occurs. To test this idea we shall need data as to whether the intervening sequences become much longer as the "C" value rises.

G. BERNARDI: I would like to stress two points. The first diagram I showed ends at man for a very simple reason, namely that I took the minimal value for each order, as had already been done by Britten. In fact, within certain orders like amphibia, you have a fantastic spread of genome sizes whereas you don't have them in other orders, like mammals or birds. The other point concerns the significance of interspersed repetitive sequences in eukaryotic genomes. This is the main discovery of renaturation kinetics. It's a pity that Britten and Davidson put so much emphasis on the regulatory role, for which there is not the least evidence, whereas the sequence may really play a role in that phenomenon of increase in genome size, which is so typical of eukaryotes and which doesn't exist in prokaryotes. In fact, all prokaryotes are within a factor of five at most in terms of genome sizes. Clearly there are two phases of evolution which can be distinguished, one in which evolution has taken place with increasing genome size and a longer one in which this has not occurred.

W.F. BODMER: Just a quick answer on the question of recombination and intervening sequences. I find that very implausible - one has to ask which came first, the chicken or the egg? I would think that the separate domains evolved and then were put together by the regions in between rather than the other way round. Recombination in higher eukaryotes is an extremely rare event at the DNA level. In higher organisms on average you've got about one crossover per chromosome per meosis which is an incredibly low frequency of recombination in terms of amounts of DNA. And, if recombination frequency were that important, you surely could easily adjust it by other ways. As in the case of mutation rates, recombination frequencies must have been carefully adjusted by recombination and there must be enormous scope for this without having to put increased amounts of DNA there to get increased recombination frequences.

G. BERNARDI: In agreement with what you say, one shouldn't forget that the intervening sequences are single copy sequences, which do not exist elsewhere in the genome. In this they are very different from the interspersed repetitive sequences which may have a role in recombination.