The Human Genome Project: a plea for basic science

Giorgio Bernardi

Basic science not only offers the fastest and deepest approach to understanding the human genome in its structure, function and evolution, but can also help the Human Genome Project in defining its priorities and strategies. This point is illustrated by research on the vertebrate genome as carried in the author's laboratory.

Although the Human Genome Project has barely taken off, it has been talked about so much over the past few years that, at present, it is probably the most widely known project in Biology. As far as one can see, there are three major reasons for such popularity: (i) the 'largest biological project ever contemplated' (as it is called in the advertising of human genome conferences) is most impressive; indeed, the project is supposed to lead to knowledge of the full sequence of the over 3 Gb (Gigabases, or billions of base pairs) of DNA that make up the human genome; (ii) it embodies a brute force approach; as such, it can be understood well beyond the narrow circle of the scientific community; (iii) the notion has been spread that, once the project is completed, everything will be known about man and his diseases. These points deserve some comment.

That the project is impressive has to be admitted, even by its strongest opponents. Sequencing over 3 billion base pairs is, in fact, such a formidable task that it was recently put aside for five years, hoping that dramatic improvements in techniques would occur in the meantime. Therefore, for the next five years, the Human Genome Project will basically be a mapping project.

As far as *physical mapping* is concerned, it is presently intended to put physical markers (recognizable sequences, like genes or anonymous single-copy sequences) at distances of 100 kb or so (1 kb, or kilobase, is one thousand base pairs) along the one-metre thread of DNA which runs through the 22 autosomes (chromosomes 1 to 22) and the 2 sex chromosomes (X and Y) of the human karyotype. Putting on a map over

Giorgio Bernardi, born in Genoa (Italy), studied Medicine in Padua and Physics in Pavia. After several years at those universities, he moved to France in 1956, where he has since been working (except for two years in Canada), first in Strasbourg and then in Paris. For the past 21 years, Dr Bernardi has been at the Laboratory of Molecular Genetics of the Jacques Monod Institute, Paris. The research interests of Dr Bernardi have been centred for many years on genome organization and evolution.

Dr Bernardi may be contacted at the following address: Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 place Jussieu, 75005 Paris, France.

30,000 landmarks is still a huge task, requiring the collaborative work of many laboratories around the world. The technology for establishing the physical map is essentially available; moreover, a number of new techniques are likely to be developed in order to proceed at a faster pace. Basically, what is needed is a series of overlapping DNA segments, which may be cloned in a number or ways, as yeast artificial chromosomes (YACs), as cosmids, or as plasmids. Particular unique sequences will then be placed on these ordered arrays of segments.

Genetic mapping is a much more laborious enterprise and its resolution (hopefully 1 centimorgan) will be about 10 times lower than that planned for the physical mapping.

The fact that the human genome project is a brute force approach raises the question as to whether intelligent approaches cannot be used to develop short-cuts in order to attain important goals, such as gene localization. One should not forget that the human genome is formed of over 90% non-coding, intergenic sequences (considered as '*junk DNA*' by certain biologists), and that there is a need to circumvent the problem of investing a lot of sequencing effort on those regions, at least in a first instance.

A number of statements have been made pushing the idea that knowledge of the primary structure of the human genome is equivalent to knowing everything about a human being. No notion could be more misleading, as will be seen from the following examples. We know the sequence of mitochondrial genomes from over 20 different species of eukaryotes. In spite of this, we do not yet understand many functional aspects of this small genome (16 kb in size in animal cells) and of its interaction with the nuclear genome. Likewise, we know a large number of sequenced viral genomes, and yet we are still very far from understanding many basic problems in virology. Such examples could be multiplied to include plasmids, as well as the genomes of chloroplasts. These remarks simply serve to stress the fact that the availability of a genome sequence is a necessary, yet far from sufficient, condition for understanding the underlying biological problems.

The points made in the previous two paragraphs suggest that we cannot hope to understand what we would like to about the human genome simply by knowing its map and, later on, its sequence, although such a knowledge will definitely expand the number of known human genes and clarify a number of genetic diseases. If the goal is understanding human genetics and the role of various genes in health and disease, the effort made on the Human Genome Project should be paralleled by a much larger one on the basic molecular genetics of man and other organisms.

What will be attempted here is to show that basic science is not only conducive to a deeper understanding of the human genome than biotechnology, but also to the development of new working hypotheses and experimental approaches having a direct impact on the strategies to follow and the priorities to choose in the Human Genome Project. This will be done by using, as an example, work carried out in the author's laboratory. But first, let us discuss the history of the word genome and its meaning.

A brief history of the word genome

The word *genome* was coined in 1920 by Winkler (11 years after the word *gene* was created by Johanssen) to denote the sum total of the genes of an organism, but only began to spread several decades later, because for many years it was not found useful by either geneticists or cytogeneticists, the former being interested in genes, the latter in chromosomes.

In 1948, it was discovered that the amount of DNA per cell is the same in all diploid cell types of a given species from multicellular eukaryotes, whereas the amount is half in gametes. The constant amount of DNA in a haploid cell of a given species was called *c-value*, or *genome size*. The constancy of genome size within a species was, however, accompanied by large variations within orders, families and genera. Such variations were unlikely to be associated with differences in organismic complexity, nor with the presumed numbers of genes, thus raising the *c-value paradox*. Moreover, the genome size of most eukaryotes obviously was in large excess relative to the likely numbers of genes. Neglecting polyploidy, which was the main factor in a number of cases, the c-value paradox and the large-size eukaryotic genomes were explained by two not mutually exclusive reasons: gene duplication, and the expansion–contraction phenomena affecting intergenic sequences. Both explanations were correct. In the evolution of eukaryotic genomes, gene duplication is common, and intergenic sequences do contract and expand. Since intergenic sequences represent the vast majority of the genome of multicellular eukaryotes (as we said, over 90% in mammals), the expansion–contraction phenomena are quantitatively by far the most important factor.

A much wider spreading of the word genome in the 1960s and 1970s was due to investigations on the sequence organization of eukaryotic genomes, using methods such as density gradient centrifugation and reassociation kinetics. These pioneering efforts were followed, after the advent of recombinant DNA technology, by an explosive development of molecular genetics. This expectedly led to an exponential increase in our knowledge of genes, from both prokaryotes and eukaryotes, including the unforeseen discovery of introns* in the latter. More recently, the use of restriction enzymes cutting at rare sequences and the separation of large DNA fragments by pulsed-field gel electrophoresis paved the way to long-range physical mapping, which became an indispensable adjunct to genetical mapping. While all these developments led to an increasing use of the word genome, what recently made it a household word was the Human Genome Project.

The meaning of the word genome

Over a lifetime of 70 years, the word genome has grown from total obscurity to enormous popularity. This success was not paralleled, however, by an increased understanding of the genome itself. Indeed, the genome still seems to be largely perceived as just the sum total of the genes of an organism (and of the intergenic sequences, if present, as in the case of eukaryotes). This view was called the *bean bag model* of the genome. According to it, the genome is just a collection of genes, which, in eukaryotes, are randomly scattered over the vast expanses of junk DNA, essentially formed by intergenic sequences.

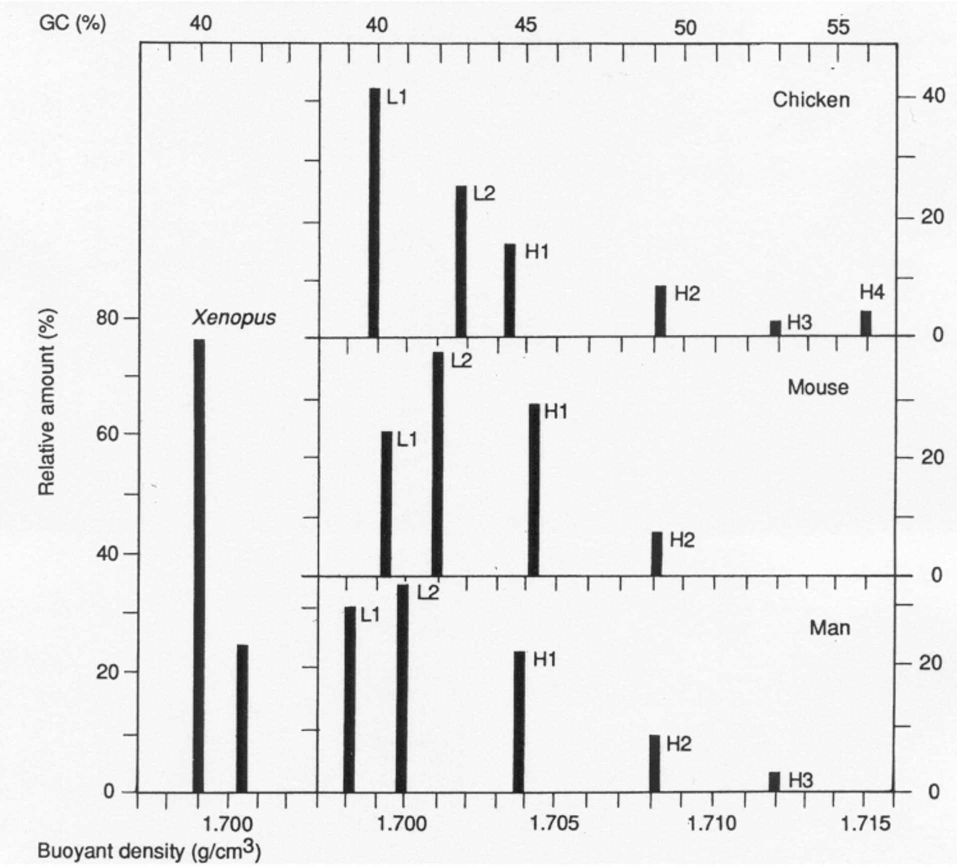
Evidence from our laboratory, however, has shown that the genome is much more than that. Indeed, the genome is an integrated structural, functional and evolutionary system that obeys precise rules, which amount to a genomic code. This view arose from investigations on the compositional properties of vertebrate genomes. The tacit assumption for this work was that base composition is a very important property of the genome.

* Intragenic sequences which are non-coding.

Although this approach was already extensively applied to the study of the mitochondrial genome of yeast and of other nuclear genomes, the discussion here will be centred on the vertebrate genome, largely using the human genome as an example.

Compositional patterns and the genome phenotype

DNA preparations from warm-blooded vertebrates (mammals and birds) are made up of large fragments (usually having a size around 100 kb) that can be separated into a



Histograms showing the relative amounts and modal buoyant densities in CsCl and GC levels of the major DNA components from *Xenopus* (the African Clawed Toad), chicken, mouse and man. Major DNA components consist of DNA fractions, obtained by density-gradient centrifugation in the presence of sequence-specific ligands, which are pooled according to their modal buoyant density in CsCl. Major DNA components derive from corresponding isochore families. The CsCl profiles of the major components are centred on their modal buoyant densities, but overlap with each other to some extent, particulary in the case of the components that are most abundant and closer in density. Satellite and minor (e.g. ribosomal DNA) components are now shown. (From Bernardi, 1989.)

number of families characterized by different GC levels (figure 1). The fragments derive, by the unavoidable mechanical and enzymatic degradation that occurs during DNA preparation, from much longer segments (over 300 kb on average) that exhibit a remarkable compositional homogeneity and which were therefore called *isochores* (for 'equal regions').

The compositional distribution of large DNA fragments, which corresponds, by far and large, to the compositional distribution of isochores, defines a *compositional pattern.* In man, GC-poor DNA fragments from isochore families L1 and L2 (figure 1)

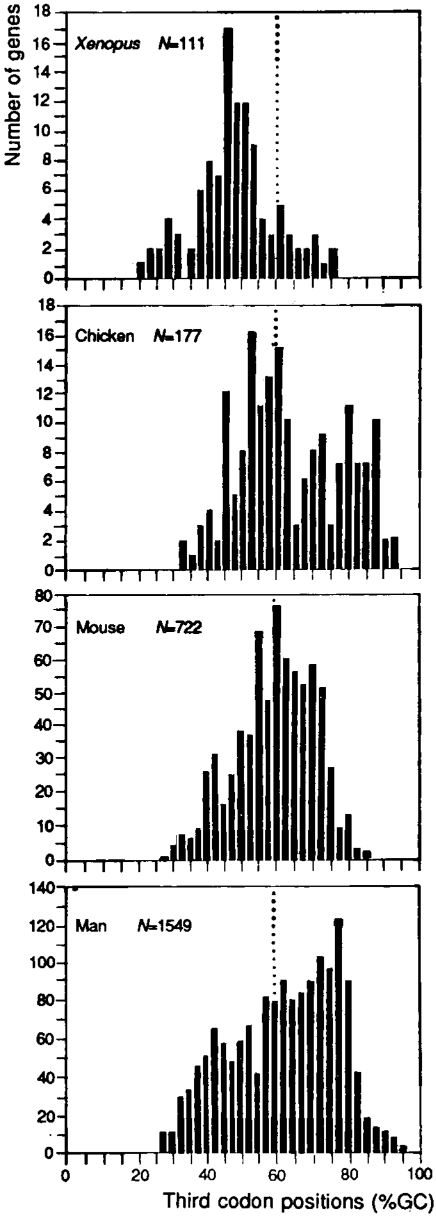


Figure 2.

Compositional distribution of third codon positions from vertebrate genes. (This distribution is the most informative because of its wider spread in composition compared to coding sequences and first or second codon positions.) The number of genes under consideration is indicated. The broken line at 60% GC is shown to provide a reference.

represent about two-thirds of the genome; GC-rich DNA fragments from isochore families H1, H2, H3 (figure 1) correspond to the remaining third. GC-poor and GC-rich isochores are interspersed with each other, as well as with the isochores corresponding to satellite (simple-sequence) DNAs and clustered repeated genes (like ribosomal genes), to form a *mosaic genome*. The strong compositional compartmentalization of the human genome (and of the genome of warm-blooded vertebrates in general) is in sharp contrast with the weak compositional compartmentalization of the genome of cold-blooded vertebrates (see figure 1), which is made up of isochores that cover a narrower GC range, and with the even narrower, unimodal compositional patterns of most invertebrates, lower (unicellular) eukaryotes and prokaryotes.

Other compositional patterns are represented by the compositional distributions of exons (or of their individual codon positions) and introns. In the genome of warm-blooded vertebrates, the compositional distribution of exons is skewed towards high GC values (figure 2), in contrast with that of DNA fragments which is skewed towards low GC values (figure 1). By contrast, genes (as well as large DNA fragments; see above and figure 1) from cold-blooded vertebrates exhibit a more symmetrical and narrower compositional distribution (figure 2).

The compositional patterns corresponding to the compositional distributions (figures 1 and 2) of large DNA fragments (or of isochores), of exons (or of individual codon positions) and introns define, in fact, *genome phenotypes*.

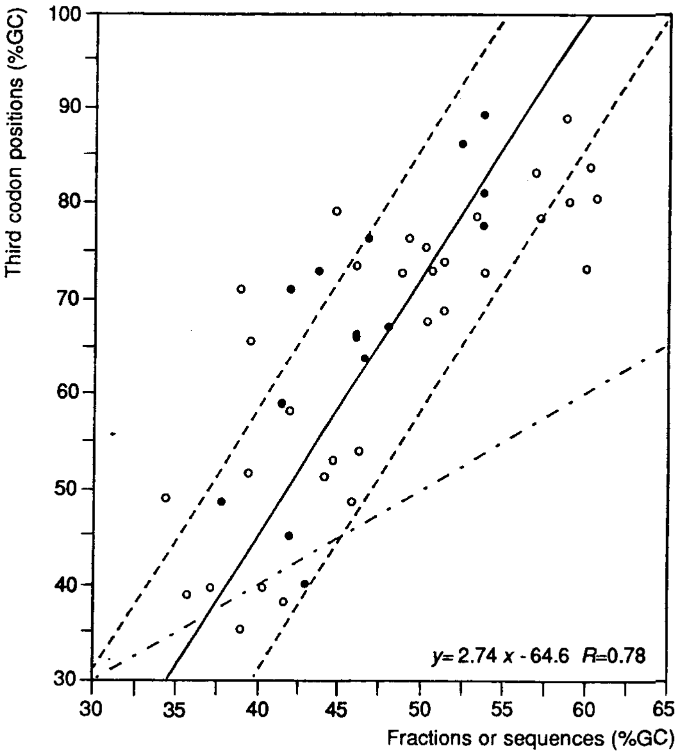


Figure 3.

Plot of GC levels of human coding sequences against the GC levels of DNA fractions or of extended, sequenced, DNA segments in which they were localized. (From Mouchiroud, D., Aïssani, B., D'Onofrio, G., Macaya, G., Gautier, C. & Bernardi, G., 1991, *Gene*, in press.)

Compositional correlations and the genomic code

The physical separation of large DNA fragments characterized by different GC levels, provided an opportunity to localize genes, or other sequences, in different isochore families by hybridization of appropriate probes. In turn, this led to the discovery that gene distribution is strikingly non-uniform in the human genome, the highest gene concentration being found in the GC-richest isochores, the H3 family, which represents less than 5% of the genome. The gene concentration in the H3 family can be estimated as at least 20 times higher than that in the L1 and L2 families.

A second discovery associated with the localization of genes in DNA fractions was that the GC levels of exons and introns are linearly correlated with the GC levels of the DNA fragments in which the corresponding genes are embedded (figure 3). Since DNA fragments are made by more than 90% intergenic sequences, the relationship links, in fact, coding and intergenic sequences. Correlations basically similar to those described for man apply to all vertebrates, whereas in the case of plants, the basic features are the same, but with quantitative differences.

In conclusion, the compositional correlations just discussed are very general, since they concern whole classes of living organisms, although not universal. By contrast, the correlation that links the three codon positions of all genomes (figure 4) is a universal

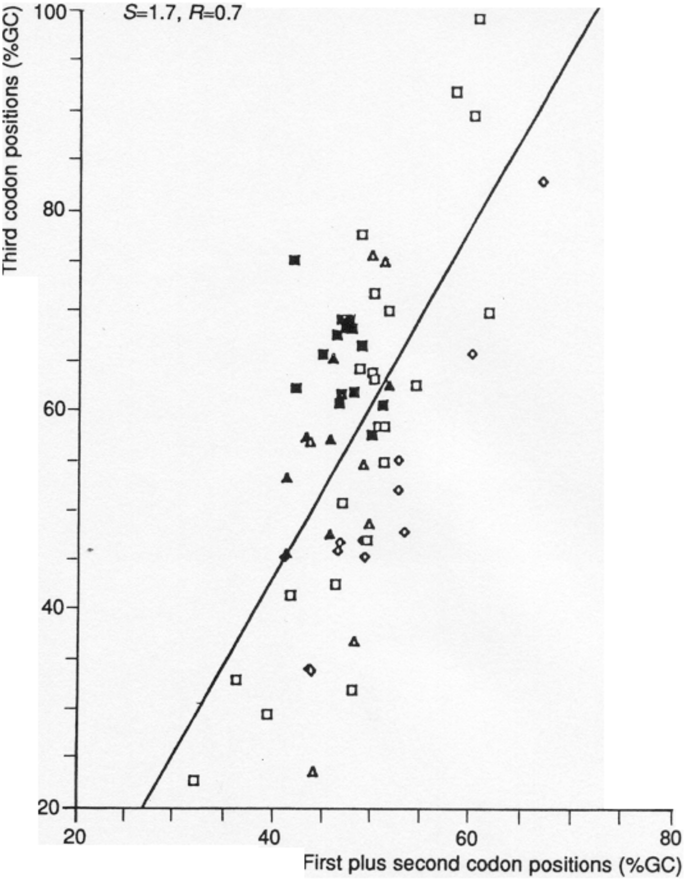


Figure 4.

Plot of GC levels of third codon positions from genes of amphibians and reptiles (▲), fishes (■), invertebrates and lower eukaryotes (△), viruses (◇) and bacteria (□) against GC levels of first + second codon positions (*S* stands for slope and *R* for the correlation coefficient). The least-square line obtained is practically coincident with that obtained for human and bacterial genes. (From Bernardi, G. & Bernardi, G., 1991, *J. Mol. Evol.*, in press.)

one. The correlations between genes and isochores and those among codon positions amount to a *genomic code*. It should be stressed that compositional patterns (or the genome phenotype) and compositional correlations (or the genomic code) are due to *compositional constraints*, acting on both coding and non-coding sequences of the genome.

Isochores, chromosomal bands and compositional mapping

Isochores are correlated with chromosomal G-bands (Giemsa-positive or Giemsa-dark bands; these are equivalent to Q-bands or Quinacrine bands) and R-bands (Reverse bands; equivalent to Giemsa-negative or Giemsa-light bands). These bands are produced by treating metaphase chromosomes with fluorescent dyes, proteases or denaturing agents. GC-poor and GC-rich isochores largely correspond to the DNA of G- and R-bands respectively, as indicated by the properties that they share (see table 1). G- and R-bands represent, therefore, a compositional pattern basically corresponding to the isochore pattern. It should be stressed that, in contrast to warm-blooded vertebrates, cold-blooded vertebrates are characterized by a poorness or absence of distinct G- and R-bands, reflecting the lower compositional comparmentalization of their genomes.

A more detailed understanding of the correlations between isochore patterns and chromosomal banding patterns can be gained through a novel approach: *compositional mapping*. Whenever long-range physical maps are available, compositional maps can be constructed by assessing GC levels around landmarks that can be probed. This simply requires the hybridization of the probes on DNA fractionated according to base composition and leads to the assessment of GC levels of DNA segments 200 kb or so in size around the sequence probed.

This approach, as applied to the long arm of human chromosome 21 (see figure 5), has provided a direct demonstration for the remarkable compositional homogeneity of G-bands and for the compositional heterogeneity of R-bands. The latter are known to correspond to different families of GC-rich isochores and also to comprise a large number of thin G-bands only visible at high resolution.

Corresponding mapping has also shown that the GC-richest isochores of the long arm of chromosome 21, which has the highest gene concentration, are located in the telomeric region. In fact, telomeres almost always correspond to R-bands, and the

Table 1

The human genome

Paleogenome (G-bands, GC-poor isochores)	Neogenome (R-bands, GC-rich isochores)
Scarcity of genes	Abundance of genes (esp. in H3)
GC-poor genes	GC-rich genes
Scarcity of CpG islands	Abundance of CpG islands
TATA box promotors	GC box promotors
Less frequent recombination	More frequent recombination
Compositional homogeneity	Compositional heterogeneity
Late replication	Early replication

Modified from Bernardi (1989); see this reference for additional details.

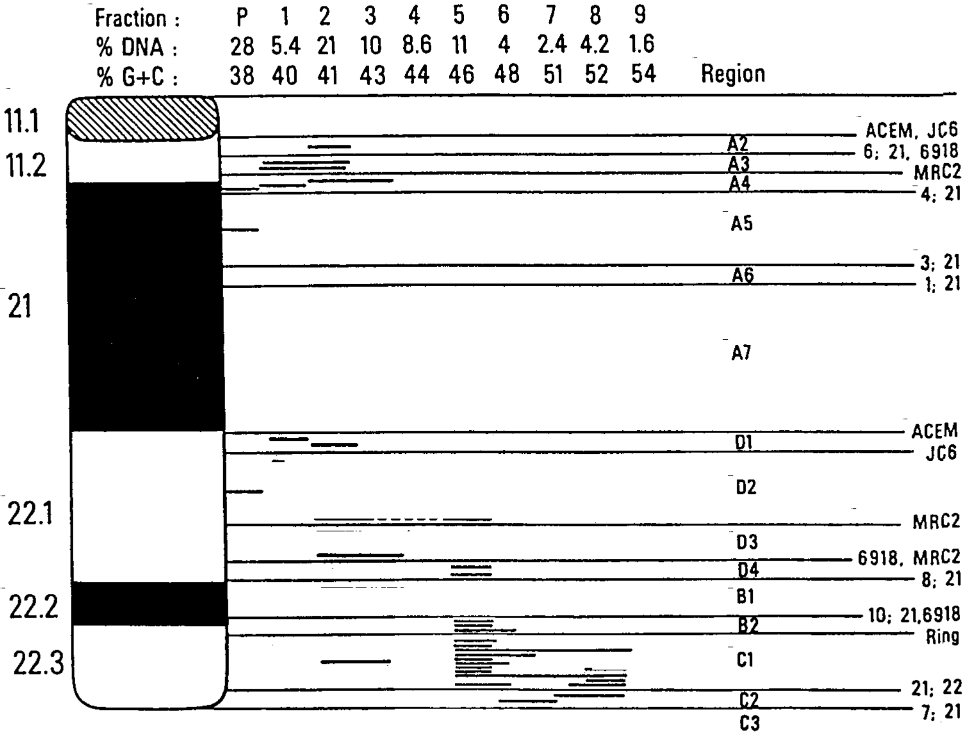


Figure 5. Compositional map of the long arm of human chromosome 21. Long horizontal lines indicate positions of the breakpoints defining chromosome segments. Short horizontal bars indicate the DNA fractions in which sequences from given chromosomal segments hybridized. This provides information on the composition of DNA segments 0·2–0·3 Mb (megabases, millions of base pairs) in size. (From Gardiner, K., Aissani, B. & Bernardi, G., 1990, *EMBO J.* **9**, 1853–1858.)

terminal regions of about half of them (including that of the long arm of chromosome 21) are the most denaturation-resistant regions of human chromosomes. We have suggested, therefore, that the GC-richest isochores correspond to such telomeric regions, a point already supported by preliminary results.

The compositional mapping of human chromosome 21, carried out as it was with only 50 probes targeted at loci scattered over about 40 Mb (megabases or millions of base pairs) of DNA, can only be considered as preliminary work. As more detailed compositional maps become available, they should provide the best high-resolution banding maps of chromosomes, without the uncertainties of cytogenetics.

Isochores and genome functions

Although we are only beginning to understand the functional relevance of isochores, some important points are already established: (i) Integrated mobile and viral sequences are preferentially located in compositionally matching isochores; (ii) translocations, sister chromatid exchanges, chromosomal abnormalities, hot-spots for mitotic chiasmata and fragile sites are preferentially located at G/R band borders and within R bands, namely at compositional discontinuities; (iii) GC levels of isochores largely determine codon usage (the differential usage of synonymous codons) and the preferential utilization of half of the aminoacids (those only having GC or AT in first and second codon positions); (iv) CpG doublets, the only potential sites of DNA methylation, and 'CpG islands' (sequences which comprise GC-rich promotors), increase from GC-poor to GC-rich isochores; (v) GC-rich isochores and R-bands replicate early and condense late in the cell cycle; the opposite is true for GC-poor isochores and G-bands.

Isochores and genome evolution

The compositional patterns of vertebrates define two modes of genome evolution. In the *conservative mode*, which is predominant in mammals and birds, mutations accumulate without causing almost any compositional changes in the genome. This can be seen by investigating the compositional distribution of DNA fragments and the chromosomal banding patterns, or, in the most rigorous way, by comparing homologous coding sequences. For instance, in phylogenetically distant mammals, homologous coding sequences do not show compositional changes, not only in first and second, but also in third codon positions.

By constrast, in the *transitional mode*, large compositional changes take place, independently of geological time and sometimes over short time-intervals. Two independent, major compositional transitions occurred in vertebrate genomes, between now-extinct reptiles (therapsids) and mammals about 200 million years ago, and between other extinct reptiles (dinosaurs) and birds about 150 million years ago. These transitions are evident when looking at the compositional distribution of DNA fragments (figure 1), and of coding sequences (figure 2), including homologous coding sequences, and also at chromosomal bands. Indeed, these transitions led to the formation of a new compositional compartment in the genomes of warm-blooded vertebrates, the *neogenome* (see table 1). In this compartment, the ancestral, GC-poor isochores were changed into GC-rich isochores that contain abundant GC-rich genes, very frequently associated with 'CpG islands'. By contrast, the other compartment of the genome of warm-blooded vertebrates, the *paleogenome* (see table 1), is characterized by its similarity to what it was, and still is, in cold-blooded vertebrates. The GC-poor isochores contain relatively sparse GC-poor genes that are only rarely accompanied by 'CpG islands'.

Conclusions

The findings obtained in our laboratory were presented here in order to stress two points.

The first is that basic science can go faster and deeper than any brute force approach if the goal is to push our knowledge further. A whole series of new important concepts have been developed during the investigations discussed here: compositional constraints act on both coding and non-coding sequences of the eukaryotic genome and lead to the formation of compositional patterns (corresponding to a genome phenotype) and to compositional correlations (corresponding to a genomic code) between coding and non-coding sequences, as well as among codon positions. Compositional patterns at the DNA level have a counterpart at the chromosomal level, GC-poor, GC-rich and GC-richest isochores corresponding to the DNA segments present in G-, R- and T-bands. The functional significance of isochores is mainly stressed by the distribution of genes which are scarce in GC-poor isochores, intermediate in frequency in GC-rich isochores, very abundant in the GC-richest isochores, and by the fact that both codon usage and aminoacid composition change in genes present in these three compositional compartments of the genome. From the evolutionary viewpoint, two modes of genome evolution have been recognized: the conservative mode in which nucleotide changes occur without any concomitant compositional change, and the transitional mode, in which nucleotide changes are accompanied by compositional changes. The conservative mode predominates in the genome of warm-blooded vertebrates which show similar compositional patterns. The transitional mode was at work at the time warm-blooded vertebrates evolved from cold-blooded vertebrates and is rather widespread in cold-blooded vertebrates.

The second point is that, in spite of the fact that no goal other than furthering our knowledge motivated our work, the basic findings just outlined have led to notions which are extremely important for the Human Genome Project: (*a*) gene concentration in the genome is characterized by a compositional gradient; genes are at least 15 times more frequent in the GC-richest isochores which represent less than 5% of the genome than in the GC-poor isochores which represent over 60% of the genome; this puts a high priority on mapping and sequencing of the GC-richest isochores which appear to be located in T-bands of metaphase chromosomes; (*b*) compositional mapping defines the compositional pattern that DNA follows along the chromosome; as such it identifies, much better than cytogenetics can do, the ultimate banding pattern of chromosomes and pins down the regions characterized by the highest gene concentrations. ■

Further reading

Bernardi, G. (1989). The isochore organization of the human genome. *Ann. Rev. Genet.*, **23**, 637–661.

Mapping and sequencing the human genome. Commission on Life Sciences, National Research Council.

Mapping our genes, Genome projects: how long, how many? Office of Technology Assessment, Congress of the United States.

Watson, J. (1990). The human genome project: past, present and future. *Science*, **248**, 44–49.