# Questions About Genomes and Genome Projects

GIORGIO BERNARDI

*Giorgio Bernardi is at the Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France. (e-mail: bernardi@citi2.fr).*

After many years, one of the genome projects has shed some light on the sequence organization—beyond the order of consecutive bases—of the eukaryotic genome. A recent paper in *Nature* (Dujon, B., et al. 1994. *Nature* **369**:371-378) shows that chromosome XI from the yeast *Saccharomyces cerevisiae* is a mosaic of guanine/cytosine (GC)-rich and GC-poor regions of around 50 kilobases. Four major GC-rich peaks were seen on the left arm and one major peak (plus a minor one) on the right arm (which is much shorter). This mosaic is seen not only in silent positions (the third bases of codons), but also in whole ORFs (open reading frames; sequences that potentially code for polypeptides) and in intergenic sequences. Moreover, yeast introns are systematically lower in GC than yeast exons, as is the case in vertebrates. Yeast represents, therefore, one more example of the generality of an isochore organization in eukaryotes.

Perhaps most interestingly, gene density in yeast chromosome XI correlates with GC-richness: In GC-rich areas, there is a higher density of genes. The finding that alternating regional variations in average GC correlate with variations in local gene density generalizes to nonvertebrate eukaryotes a similar situation previously found in vertebrates. But the significance of genome composition goes beyond this: It raises important questions both about genomes and about genome projects.

To start with, we have to go back to our understanding of the term, "genome." The word was coined in 1920 by Hans Winkler to define the (haploid) ensemble of the genes of an organism, although the definition now includes noncoding regions. Surprisingly, three-quarters of a century after Winkler many biologists (including, apparently, some genome project decision-makers) still visualize the genome only according to the original, purely operational, definition. The genome is more than a bag of genes or strings of DNA. Defined in this vein, a car would be merely a means, composed largely of metal, of achieving superhuman speed on land.

The fact of the matter is that we do know a great deal more than this about the organization of the genome (for a review, see Bernardi, G. 1993. *Gene* **135**:57-66) and we ought to think about applying that knowledge in the design of genome projects. For instance, we know that the genome of vertebrates are mosaics of long (>300 Kb) compositionally homogeneous segments of DNA, called isochores, of different base compositions. In the genomes of warm-blooded vertebrates, isochore GC-content ranges from 30 to 60 percent; in that of cold-blooded vertebrates, the range is much narrower and does not reach very high GC levels.

We know also that the concentration of genes is strikingly nonuniform; in mammals, the gene concentration is at least 20 times higher in the GC-richest isochore family (representing only 4 percent of the genome) than in the GC-poor isochore families (that represent more than 60 percent of the genome). Indeed, such nonuniform gene distribution is also present in cold-blooded vertebrates. What happened at the transitions from reptiles to mammals and from reptiles to birds was that the gene-rich regions of the genomes of the reptiles underwent a GC increase (forming the neogenome of warm-blooded vertebrates) whereas the gene-poor regions did not undergo a compositional change (and formed the paleogenome of warm-blooded vertebrates).

A comparison of aligned homologous genes unambiguously demonstrated that the increase in GC was due to a process of directional fixation of mutations. This can be visualized as caused by a bias in mutations (during repair and/or replication, two phenomena associated with transcription) and/or by selection for an isochore composition that was more advantageous. In both cases, the change has a functional relevance, as also stressed by the fact that it occurred in the gene-rich, and not in the gene-poor, regions of the genome. The isochore mosaic appears, therefore, to reflect functionally meaningful compositional requirements at the coding and contiguous noncoding sequences.

Indeed, the significance of the isochore organization of the mammalian genome is not only evolutionary but also functional. The GC-richest isochores have not only the highest gene concentration (and the highest concentration of housekeeping genes), but also the highest concentration of CpG islands (nonmethylated, GC-rich, regulatory sequences located upstream of genes), an open chromatin structure, the highest transcription and recombination levels, and extreme codon usages.

These features of genome organization indicate that the genome should be regarded not only as the source of genetic information, but also as a structure whose functions are modulated by base composition (through codon usage, transcription rate, mRNA stability, etc.).

What are the implications for genome projects? If genes are more concentrated in GC-rich regions, then it might make scientific and economic sense to concentrate mapping and sequencing in those regions. Magnificent recent efforts have produced a human genome map that covers 90 percent of the genome. Unfortunately, the remaining 10 percent that, for technical reasons, has proved difficult to map, is highly GC-rich. The unmapped 10 percent may encompass 50 percent of the genes. ///