Evolutionary History of the Human Genome

Giorgio Bernardi, *Laboratory of Molecular Evolution, Stazione Zoologica Anton Dohrn, Naples, Italy*

The compositional evolution of the vertebrate genome, from fishes to humans, is of very general interest because of its bearing on the role of natural selection in the evolutionary process.

Introduction

Compositional genomics is an approach to the problem of the organization of eukaryotic genomes. Initially this approach consisted of analysing the base composition of complex eukaryotic genomes (such as the nuclear genomes of vertebrates) by using density gradient centrifugation in the presence of sequence-specific ligands. This approach revealed that vertebrate genomes are mosaics of very long stretches of deoxyribonucleic acid (DNA) with fairly homogeneous GC levels (the percentage of guanine + cytosine), the isochores, which belong to a small number of families characterized by different base compositions. This compositional approach was extended to nucleotide sequences when these became available. When comparative studies were made of the organization of the nuclear genomes of vertebrates, compositional differences were found and were shown to have an evolutionary relevance. **See also**: Genome Organization of Vertebrates; Isochores

Major Shifts

DNAs from all warm-blooded vertebrates exhibit high compositional heterogeneities and strongly asymmetrical CsCl bands (the bands formed by centrifuging DNA to equilibrium in a CsCl density gradient), whereas DNAs from cold-blooded vertebrates are generally characterized by low compositional heterogeneities and by only slightly asymmetrical CsCl bands. Both differences, in heterogeneity and asymmetry, are due to the presence in the genomes of warm-blooded vertebrates of a small percentage (approximately 15%) of GC-rich DNA molecules that are less GC-rich in the genomes of most cold-blooded vertebrates (**Figure 1a**). These differences can also be seen at the level of coding sequences. Indeed, coding sequences from cold-blooded vertebrates are relatively homogeneous in base composition and are generally characterized by low GC levels, whereas coding sequences from warm-blooded vertebrates are compositionally much more heterogeneous and reach very high GC levels, up to 100% GC in the third codon positions of genes (**Figure 1b**). The similarity of the two patterns (a and b of **Figure 1**) is due to the fact that coding sequences are compositionally correlated with the isochores that embed them.

Since the genomes of mammals and birds derive from those of two separate lines of ancestral reptiles (see **Figure 2**), these findings indicate that two major compositional changes occurred independently in the distinct ancestral lines leading to warm-blooded vertebrates. In other words, two major intragenomic shifts took place in the genomes of the ancestors of present-day mammals and birds.

Compositional Evolution: Genome Phenotypes, the Transitional and the Conservative Modes of Evolution

In vertebrates, the compositional patterns of isochores and of coding sequences, as well as the GC_3 levels of orthologous coding sequences, were found to be strikingly different between fish/amphibians on the one hand and mammals/birds on the other, reptiles showing intermediate patterns. More precisely, at the transition between cold- and warm-blooded vertebrates, the moderately GC-rich ancestral genome core of cold-blooded vertebrates underwent a remarkable GC increase to become the genome core of mammals and birds (see **Figure 3**). This implied the existence of a *transitional* (or *shifting*) *mode* of evolution that took place independently in the ancestral genomes of mammals and birds, and yet led to similar isochore patterns that were maintained in mammalian and avian genomes. In contrast, the genome desert did not undergo any major compositional change, except for the decrease of mC (methylcytosine) and CpG which affects the whole genome (see later).
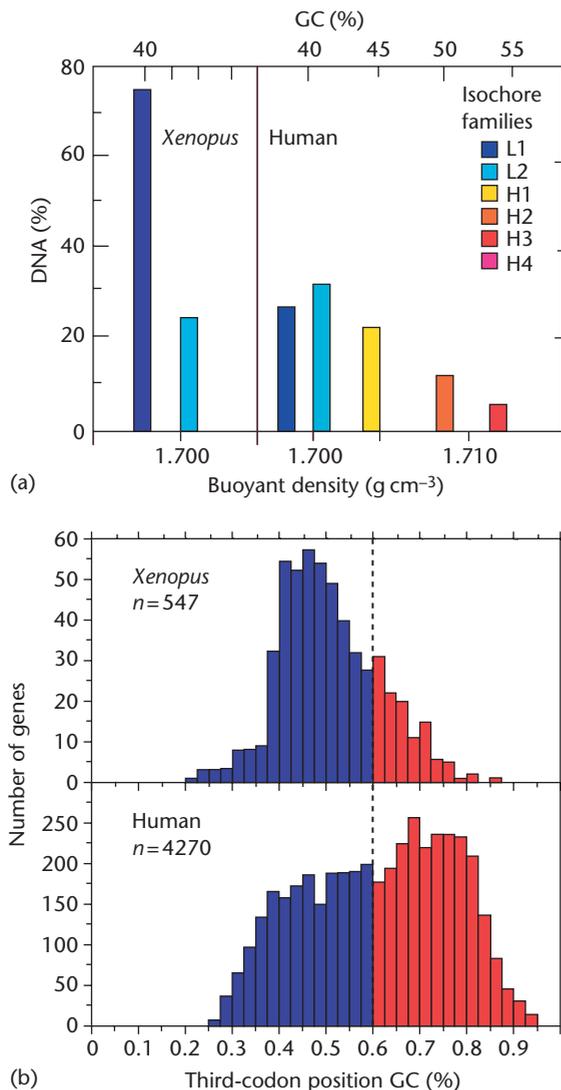
(a)

(b)

Figure 1 (a) Compositional pattern of vertebrate genomes. Histograms showing the relative amounts, modal buoyant densities and modal GC levels of the major DNA components (the families of DNA fragments derived from different isochore families) from *Xenopus* and human, as estimated after fractionation of DNA by preparative density gradient in the presence of a sequence-specific DNA ligand (Ag$^+$ or bis(acetatomercurimethyl)dioxane, BAMD). Satellite and minor DNA components (such as ribosomal DNA) are not shown. (b) Compositional patterns of coding sequences represented by GC$_3$ values from *Xenopus* and human. The broken line at 60% GC$_3$ is shown as a reference. Modified from Bernardi G (1995) The human genome: organization and evolutionary history. *Annual Reviews of Genetics* **29**: 445–476.

Figure 2 The maximum likelihood phylogeny of combined sequences from 11 proteins, 1943 amino acids (from Hedges SB and Poling LL (1999) A molecular phylogeny of reptiles. *Science* **283**: 998–1001), is compared with the CsCl profiles (from Thiery JP, Macaya G and Bernardi G (1976) An analysis of eukaryotic genomes by density gradient centrifugation. *Journal of Molecular Biology* **108**: 219–235). Red frames concern warm-blooded vertebrates (birds, primates), orange and blue frames cold-blooded vertebrates (reptiles showing a different degree of compositional heterogeneity). The scale bar indicates amino acid substitutions per site. From Cruveiller S, D'Onofrio G and Bernardi G (2000) The compositional transition between the genomes of cold- and warm-blooded vertebrates: codon frequencies in orthologous genes. *Gene* **261**: 71–83.

Compositional differences among genomes from vertebrates belonging to the same classes (fishes, amphibians, etc.) could, however, be observed as the result of 'whole genome (or horizontal) shifts'. These shifts apparently are the consequence of changes in genome size, in environmental conditions, in mutation rates and in the intensity and duration of the mutational AT bias, i.e. the predominance of GC→AT versus AT→GC (A, adenine; G, guanine; C, cytosine and T, thymine) changes. This bias is
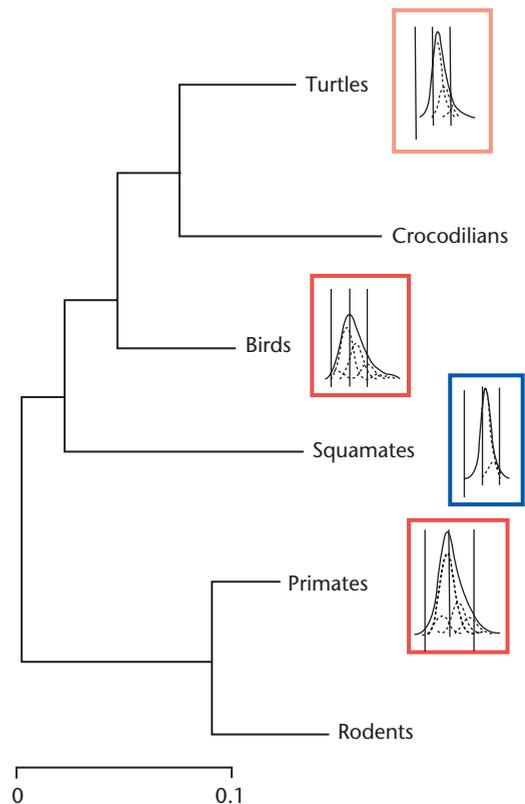
likely to be due, as in the case of *mutator mutations* of bacteria and the archaea, to mutations in the genes coding for subunits of the replication machinery.

The compositional transition between cold- and warm-blooded vertebrates could not be accounted for by the random nucleotide substitutions of the neutral theory (see also later) and required, therefore, another explanation. The *thermodynamic stability hypothesis* proposed that the GC increase of the genome core accompanied the emergence of homoeothermy and simultaneously provided the selective advantages of increased stabilities of DNA, RNA (ribonucleic acid) and proteins. Indeed, high GC levels in coding sequences not only increase the stability of DNA and of the stem structures of transcribed RNAs, as demonstrated by the GC increases of corresponding stems of ribosomal RNAs from cold- to warm-blooded vertebrates, but also favour amino acids that thermodynamically stabilize
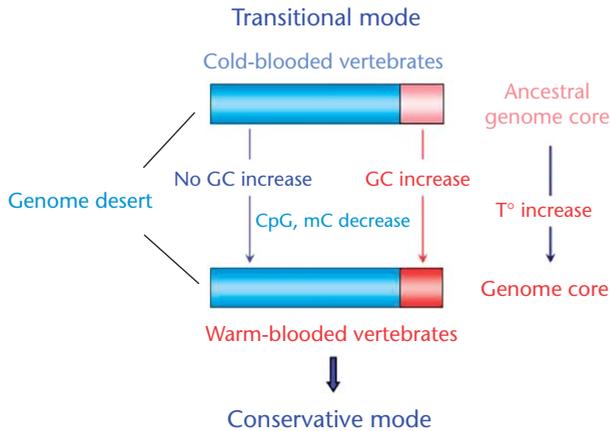
Figure 3 Scheme of the compositional evolution of vertebrate genomes. At the transition from cold- to warm-blooded vertebrates, the gene-dense, moderately GC-rich *ancestral genome core* (pink box) became the gene-dense, GC-rich *genome core* (red box), whereas the GC-poor and gene-poor (blue box) *genome desert* did not undergo any major compositional change. This transitional (or shifting) mode, which was accompanied by an overall decrease of CpG doublets and mC (methylcytosine), was followed by a conservative mode of genome evolution in which compositional patterns were maintained. Reproduced from Bernardi G (2007) The neo-selectionist theory of genome evolution. *Proceedings of the National Academy of Sciences of the USA* **104**: 8385–8390.

Figure 4 An example of isochore conservation between syntenic chromosome regions of dog and human. Reproduced from Bernardi G (2007) The neo-selectionist theory of genome evolution. *Proceedings of the National Academy of Sciences of the USA* **104**: 8385–8390.

proteins, as indicated by observations of Argos *et al.* and Zuber that were recently confirmed by Nishio *et al.*

This selectionist explanation was supported (i) by the similar genome changes occurring in the independent lines of mammals and birds; (ii) by the decreases of CpG doublets and mC, which are correlated with increasing body temperatures in vertebrates ranging from Antarctic fishes to mammals; (iii) by the variable compositional heterogeneity and methylation levels (intermediate between those of fishes/amphibians and mammals/birds) of the genomes from reptiles, which are known to have different body temperatures and thermal regulations; (iv) by the mammalian-like isochore organization of the genomes of subtropical and tropical insects (*Drosophila*, *Anopheles*) and (v) by the increase of GC levels that accompanies the increase in optimal growth temperatures in many families of bacteria and the archaea.

The thermodynamic stability hypothesis stressed the moulding effect of the environment on the genome. In fact, our use of the old-fashioned distinction between cold- and warm-blooded vertebrates was meant to suggest that the cause of genome changes was *body temperature*. The hypothesis could also explain why, at the emergence of a stable body temperature around 40°C (homoeothermy), the *open chromatin* of the genome core, which has an expanded configuration at the centre of the interphase nucleus, required a GC increase to be stabilized, whereas the *closed chromatin* of the GC-poor genome desert did not, being stabilized by its own compact structure packed against the nuclear membrane. The different chromatin structure could also explain why retroviral sequences initially integrate in GC-rich regions.
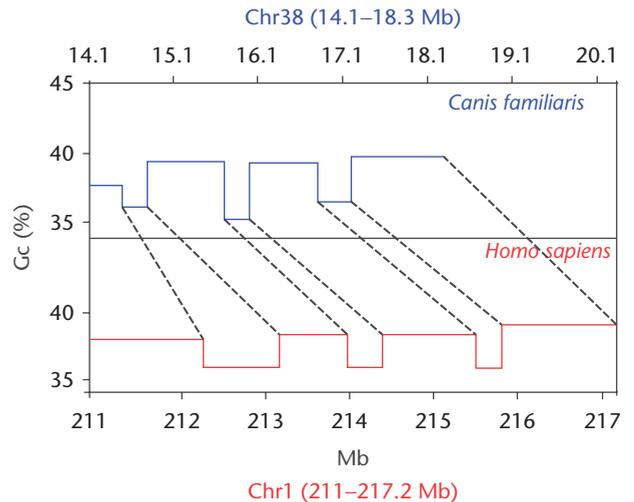
The transitional mode of evolution stopped in mammals after they emerged from their common ancestor, suggesting that an equilibrium was reached. The evidence for this compositionally *conservative mode* of evolution was initially provided by the similarity of the compositional distribution of coding sequences and of GC_3 levels of orthologous coding sequences of mammals from different orders. Recent results showed that, in addition, the patterns of orthologous (or syntenic) isochores are conserved in different mammalian orders (see e.g. **Figure 4**). This is an astonishing result in view of the mutational AT bias of vertebrates and of the fact that the species under consideration have been diverging for 90 million years. Since the neutral substitution rate has been estimated as 3.9×10^{-9} substitutions per year, approximately 70% nucleotide substitutions took place. This would approximately correspond to 54% sequence identity if one accounts for multiple and parallel substitutions. Some differences in size and in GC levels between syntenic isochores do exist, but they are largely accounted for by inserted mobile elements, and by the whole-genome shifts mentioned earlier. Needless to say, the remarkable conservation of both isochore size and average base composition of isochore families in all vertebrates adds two very important properties to the general picture just presented.

The conservative mode of evolution was originally explained by 'negative selection acting at a regional (isochore) level to eliminate any strong deviation from presumably functionally optimal composition of isochores'. It should be stressed that the compositional conservation of syntenic isochores (see **Figure 3**) not only reinforces the original idea of negative selection, but also implies a functional role of noncoding sequences in the genome, which is also suggested by a number of other considerations. Objections to selection and alternative

hypotheses (in particular the biased gene conversion) for the formation and maintenance of the GC-rich isochores of mammals and birds have been discussed in detail elsewhere and shown not to hold.

In conclusion, selection appears to account for both the formation and the maintenance of isochores, two phenomena that cannot be explained by the neutral theory, which, incidentally, was elaborated without any knowledge of isochore patterns. In fact, comparisons of orthologous mammalian genes or proteins (such as those used at that time) unfortunately missed not only the compositional heterogeneity of mammalian genes within the same genome, but also the compositional changes that took place between cold- and warm-blooded vertebrates. However, while a random fixation of neutral mutants (or a biased genome conversion) could not generate GC-rich isochores and would erode the isochore pattern over evolutionary time, the majority of mutations per se could only be neutral and nearly neutral. Indeed, in genomes in which noncoding sequences may represent 99% of DNA, no single-nucleotide change could have a strong enough selective advantage or disadvantage, at least in the vast majority of cases. There was, therefore, a need to reconcile the neutrality or near-neutrality of the point mutation process with selection at the regional level. This was done by the neoselectionist theory of genome evolution. **See also**: Chromatin Structure and Domains; Chromosomal Bands and Sequence Features; Chromosomes 21 and 22: Comparisons; Chromosomes 21 and 22: Gene Density; Codon Usage; CpG Islands and DNA Methylation; DNA Methylation: Evolution; Evolution: Neutralist View; Evolution: Selectionist View; GC-rich Isochores in the Interphase Nucleus; Gene Structure and Organization; Genes: Types; Genome Organization of Vertebrates; Isochores; L Isochore Map: Gene-poor Isochores; Long Interspersed Nuclear Elements (LINEs); Mitochondrial Genome: Evolution; Mutational Change in Evolution; Orthologues, Paralogues and Xenologues in Human and Other Genomes; Polyploid Origin of the Human Genome; Promoters: Evolution; Protein Coding; Selective and Structural Constraints; Transposons

Further Reading

Bernardi G (1995) The human genome: organization and evolutionary history. *Annual Reviews of Genetics* **29**: 445–476.

Bernardi G (2000) Isochores and the evolutionary genomics of vertebrates. *Gene* **241**: 3–17.

Bernardi G (2000) The compositional evolution of vertebrate genomes. *Gene* **259**: 31–43.

Bernardi G (2004, reprinted in 2005) *Structural and Evolutionary Genomics. Natural Selection in Genome Evolution*. Amsterdam: Elsevier.

Bernardi G (2007) The neo-selectionist theory of genome evolution. *Proceedings of the National Academy of Sciences of the USA* **104**: 8385–8390.

Carroll RL (1997) *Patterns and Processes of Vertebrate Evolution*. Cambridge: Cambridge University Press.

Cruveiller S, D'Onofrio G and Bernardi G (2000) The compositional transition between the genomes of cold- and warm-blooded vertebrates: codon frequencies in orthologous genes. *Gene* **261**: 71–83.

Graur D and Li WH (1999) *Fundamentals of Molecular Evolution*, 2nd edn. Sunderland, MA: Sinauer Associates.

Hedges SB and Poling LL (1999) A molecular phylogeny of reptiles. *Science* **283**: 998–1001.

Saccone S, Federico C, Andreozzi L, D'Antoni S and Bernardi G (2002) Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. *Gene* **300**: 169–178.

Saccone S, Pavliček A, Federico C, Paçes J and Bernardi G (2001) Genes, isochores and bands in human chromosomes 21 and 22. *Chromosome Research* **9**: 533–539.

Thiery JP, Macaya G and Bernardi G (1976) An analysis of eukaryotic genomes by density gradient centrifugation. *Journal of Molecular Biology* **108**: 219–235.