Structural and Evolutionary Genomics

Natural Selection in Genome Evolution

Giorgio Bernardi
Structural and Evolutionary Genomics

Natural Selection in Genome Evolution

GIORGIO BERNARDI

Stazione Zoologica Anton Dohrn
Naples, Italy
Preface

The main purpose of this book is to present our investigations in the areas of structural and evolutionary genomics, to critically review the relevant literature and to draw some general conclusions. Even if "functional genomics" is not included in the title, a number of functional implications derived from structural and evolutionary genomics will be discussed. While the majority of the book concerns genome organization, the last Parts present "a long argument" on the role of natural selection, "the preservation of favourable variations and the rejection of injurious variations" (Darwin, 1859), in genome evolution.

I intended to write this book for several years, but I hesitated mainly because firm conclusions on the role of natural selection in genome evolution had not yet been reached. Even if new results may modify the picture presented here, I now feel that its main features are correct, and that the time is ripe for publishing this overview.

Basically, the book presents experimental and conceptual advances in two major areas. The first one is genome organization. In spite of recent spectacular progress in genome sequencing, the remark that "a large amount of detail is available, but comprehensive rules about the organization of genome have not yet emerged" (Singer and Berg, 1991) still applies to the current literature. Our main discoveries, concerning the compositional compartmentalization of the vertebrate genome into a mosaic of isochores, the genome phenotypes, the genomic code, the bimodal distribution of genes and its correlation with functional properties, have led for the first time to a unified view of the eukaryotic genome as an integrated ensemble.

The second area is genome evolution. Our findings could not be accounted for by any of the current molecular evolution theories, since they were all based on single-nucleotide changes, and did not (and could not) take into consideration regional and compositional changes. We have been able to build a model of genome evolution, the neo-selectionist model, which accommodates not only some key features of the classical selection theory (essentially the selection of single-nucleotide changes in coding and regulatory sequences), but also those of the neutral theory (basically the random fixation of selectively neutral or nearly neutral changes in noncoding sequences). The neutral and nearly neutral changes certainly represent the majority of the changes in genome evolution, but they are finally controlled at the regional level by natural selection (essentially negative selection). In other words, the neo-selectionist model puts the neutral view of the genome into a new selectionist frame.

The book starts (Part 1) with a short history of the different views concerning the genome, a brief narrative of our early investigations, and a discussion of the molecular approaches that we used. Part 2 deals with a small model genome, the mitochondrial genome of yeast, which shed light on the large genome in the nucleus.

In the central section of the book, Parts 3 and 4 outline the compositional properties of the vertebrate genome, namely the compositional patterns of DNA molecules and of coding sequences, as well as the compositional correlations between coding and non-coding sequences, whereas Parts 5, 6 and 7 discuss the most important properties of the vertebrate genome: the distributions of genes, of transposons and of integrated viral
sequences in the genome and in chromosomes. This book is, however, not limited to the vertebrate genome, but also concerns other eukaryotic genomes, in particular plant genomes, as well as prokaryotic genomes (Parts 8 and 9).

The book ends with Part 10, which examines the correlations between gene composition and protein structure, Part 11, which considers how the organization of the vertebrate genome evolved in time, and Part 12, which discusses the general causes and mechanisms of this evolution. A recapitulation and our conclusions concerning the relative roles of natural selection and random drift in the evolution of living organisms are presented in the final sections.


The names and the contributions of the many people who participated in the investigations described in this book can be gathered from the references. I would like, however, to mention the names of those who either played a particularly important role in some phases of this work, or did more than the references suggest.

The first group comprises several people. My brother Alberto closely collaborated with me both in Strasbourg in the 1960’s, on the preparation of DNases, exonucleases, phosphatases etc., which had never been prepared before, and later in Paris. In the early 1970’s, Jean-Paul Thiéry, Gabriel Macaya and Jan Filipski set the foundations for the investigations that kept us busy for many years, while Dusko Ehrlich was the major contributor to our approach on the frequency of oligonucleotides in DNAs. My second son, Gregorio, started the computer analysis of DNA sequences in 1980, with the help of Jacques Ninio. My youngest son, Giacomo, initiated our investigations in molecular evolution in 1985 and has been collaborating with me since then. In the 1990’s, Giuseppe D’Onofrio was responsible for pursuing further our investigations on both the organization and the evolution of the mammalian genome together with Simone Caccio, Oliver Clay, Kamel Jabbari, Dominique Mouchiroud, Hector Musto and Serguei Zoubak. In more recent years and until present, Giuseppe D’Onofrio, Oliver Clay, Kamel Jabbari and Héctor Musto were joined by Fernando Alvarez-Valin, Nicolas Carels, Stéphane Cruveiller and Adam Pavliček. Salvo Saccone was behind all the cytogenetic work in which compositional DNA fractions were used for in situ hybridization. Along the yeast mitochondrial research line, the major contributions came from Giuseppe Baldacci, Miklos de Zamaroczy, Godeleine Faugeron-Fonty, Regina Goursot, Gianni Piperno, Ariel Prunell, and Edda Rayko. The second group comprises Claude Cordonnier, Anne Devillers-Thiery, Audrey Haschemeyer and Alla Rynditch, who made investigations on hydroxyapatite chromatography, oligonucleotide frequencies, fish genomics and retroviral integrations, respectively. I certainly do not forget my faithful technicians Andréa Silvert and Henri Stebler, my draftman/photographer Philippe Breton, and Martine Brient, my secretary for almost thirty years.

I also wish to thank Fernando Alvarez-Valin, Giacomo Bernardi, Giuseppe D’Onofrio,
Regina Goursot, Kamel Jabbari, Adam Pavliček, Edda Rayko, and, especially, Oliver Clay and Héctor Musto for critical reading of sections of this book. Its preparation would have been impossible without the intelligent, competent and dedicated help of Gianna Di Gennaro and Romy Sole. I am grateful to Francisco Ayala, Takashi Gojobori, Daniel Hartl, Toshimichi Ikemura, Masatoshi Nei, Tomoko Ohta and Emile Zuckerkandl for their interest and encouragement. Last but not least, I wish to thank Dr. Arthur Koedam of Elsevier for his patience and understanding.

The first draft of this book was prepared at Hopkins Marine Biology Laboratory of Stanford University, Pacific Grove, in August 2001, thanks to the hospitality of George Somero. The book was written in the congenial atmosphere of the Stazione Zoologica Anton Dohrn, where it was completed in July 2003. Two notes were added in proof in early November 2003.

Finally, I would like to mention that some ideas presented in this book were developed during extensive travel and field work (essentially linked to specimen collection) in faraway places, often with my wife Gabriella and/or my son Giacomo. It was an honour, and a pleasure, to have, in some of these trips, the company of Professor Richard Darwin Keynes, FRS, the great grandson of Charles Darwin.

I would like to offer my sincere apologies to two groups of people. The first group comprises the colleagues whose work I am criticizing. I wish to make clear that criticisms were not just raised for polemical reasons, but because the analysis of a wrong experiment, or of a wrong viewpoint, can advance our understanding of a problem. Moreover, it is often instructive to present the background of wrong ideas against which new facts had to emerge. My feeling is that science makes progress, like evolution, more by negative selection (of wrong facts and views, which are abundant), than by positive selection (of good ideas, which are rare). Let me add that my personal opinion is that in science the principle “Amicus Plato, sed magis amica veritas” should prevail over any other consideration, diplomatic and otherwise.

The second group is that of the readers of this book. Covering over 40 years of work in one volume was not easy. For the sake of speeding up the preparation of this book, I did not hesitate to use verbatim quotations from our papers, especially the most recent ones. I hope the readers will excuse me for not having spent more time in polishing the style and smoothing out the jumps from one subject to another. They should, however, remember that this is not a textbook but a scientific monograph, that often deals with subjects at the border of our knowledge. Moreover, this book is focused on the general picture rather than on details, on the rule rather than on the exceptions. For this very reason, some subjects that are very important in themselves, were treated only in a cursory way, if their relevance to the main line of this book was marginal. I tried to be as clear as possible, while solving two problems, namely introducing methodological approaches which might not be generally familiar to the readers, and sketching a complex picture. Including all this information in the book was not a minor enterprise. This task was, however, made simpler by three factors. First, the main line of the book presents investigations carried out in a single laboratory. Second, the molecular biology approaches that we used provided results that could stand time (the buoyant density of DNA, for example, does not become obsolete over the years). Third, most of the data presented are very recent. In fact, some of the
articles referred to will still be in press at the time this book will appear. This time will coincide with the 50th anniversary of the double helix paper by Francis Crick and Jim Watson, whom I salute and greet, and the 20th anniversary of the publication of *The Neutral Theory of Molecular Evolution* by Motoo Kimura, the great scientist to whose memory I pay homage.

Giorgio Bernardi
Contents

PREFACE ...................................................................................................................................... VII

PART 1: INTRODUCTION ............................................................................................................. 1
1.1 The genome: a short history of different views ................................................................. 3
1.2 Population genetics and molecular evolution ................................................................. 4
1.3 Three remarks on terminology ....................................................................................... 5
1.4 A brief chronology of our investigations ....................................................................... 5
1.5 Molecular approaches to the study of the genome ......................................................... 10

PART 2: LESSONS FROM A SMALL DISPENSABLE GENOME.
THE MITOCHONDRIAL GENOME OF YEAST ........................................................................ 19

Chapter 1. The mitochondrial genome of yeast and the petite mutation ......................... 21
1.1 The “petite colony” mutation .......................................................................................... 21
1.2 The petite mutation is accompanied by gross alterations of mitochondrial DNA .......... 23
1.3 The AT spacers and the deletion hypothesis .................................................................. 23
1.4 The petite mutation is due to large deletions ............................................................... 25
1.5 The GC clusters ............................................................................................................ 26
1.6 The excision sites .......................................................................................................... 26
1.7 Genomes without genes ............................................................................................... 28

Chapter 2. The origins of replication .................................................................................... 31
2.1 Excision and recombination ......................................................................................... 31
2.2 The canonical and the surrogate origins of replication of petite genomes .................. 32
2.3 The replication of petite genomes and the phenomenon of suppressivity ................... 35
2.4 The ori sequences as transcription initiation sites ....................................................... 37
2.5 The effect of flanking sequences on the efficiency of replication of petite genomes ....... 38
2.6 The ori− petites 14 and 26 ......................................................................................... 39
2.7 Temperature and the replicative ability of ori− petites 14 and 26 ............................... 42

Chapter 3. The organization and evolution of the mitochondrial genome of yeast ......... 43
3.1 The organization of the mitochondrial genome of yeast ............................................ 43
3.2 The evolutionary origin of ori sequences ..................................................................... 44
3.3 The evolutionary origin of the GC clusters ................................................................. 45
3.4 The evolutionary origin of the AT spacers and the var 1 gene ................................... 45
3.5 The non-coding sequences: evolutionary origin and biological role .......................... 46

PART 3: THE ORGANIZATION OF THE VERTEBRATE GENOME ........................................ 49

Chapter 1. Isochores and isochore families ........................................................................ 51
1.1 The fractionation of the bovine genome ....................................................................... 51
1.2 The fractionation of eukaryotic main-band DNAs ...................................................... 53
1.3 Isochores and isochore families ................................................................................... 56
1.4 Isochores and the draft human genome sequence ..................................................... 63
1.5 Other misunderstandings about isochores ................................................................... 71

Chapter 2. Compositional patterns of coding sequences .................................................... 75

Chapter 3. Compositional correlations between coding and non-coding sequences ....... 77
PART 4: THE COMPOSITIONAL PATTERNS OF VERTEBRATE GENOMES

Chapter 1. The fish genomes

1.1 Compositional properties: a CsCl analysis
1.2 Compositional properties: a Cs2SO4/BAMDI analysis
1.3 Compositional properties: an analysis of long sequences
1.4 Compositional properties of coding sequences and introns
1.5 Compositional correlations

Chapter 2. Amphibian genomes

Chapter 3. Reptilian genomes

Chapter 4. Avian genomes

Chapter 5. Mammalian genomes

PART 5: SEQUENCE DISTRIBUTION IN THE VERTEBRATE GENOMES

Chapter 1. Gene distribution in the vertebrate genome

1.1 The distribution of genes in the human genome: the two gene spaces
1.2 Properties of the two gene spaces
1.3 The distribution of genes in the vertebrate genomes

Chapter 2. The distribution of CpG islands in the vertebrate genome

Chapter 3. The distribution of CpG doublets and methylation in the vertebrate genome

3.1 CpG doublets
3.2 Two different CpG levels in vertebrate genomes
3.3 Two different methylation levels in vertebrate genomes

PART 6: THE DISTRIBUTION OF INTEGRATED VIRAL SEQUENCES, TRANSPONDONS AND DUPLICATED GENES IN THE MAMMALIAN GENOME

Chapter 1. The distribution of proviruses in the mammalian genome

1.1 The integration of retroviral sequences into the mammalian genome
1.2 The bimodal compositional distribution of retroviral genomes
1.3 The localization of integrated viral sequences in the host genome
1.4 An analysis of integration sites near host cell genes
1.5 The correlation between the isochore localization of integrated retroviral sequences and their transcription
1.6 Integration in "open" chromatin and/or near CpG islands
1.7 The causes of the compartmentalized, "isopycnic" localization of viral sequences

Chapter 2. The distribution of repeated sequences in the mammalian genome

2.1 Alu and LINE repeats in human isochromes
2.2 The evolutionary origin of repeat distribution: different viewpoints
2.3 Repeated sequences in coding sequences?

Chapter 3. The distribution of duplicated genes in the human genome

PART 7: THE ORGANIZATION OF CHROMOSOMES IN VERTEBRATES

Chapter 1. Isochromes and chromosomal bands

Chapter 2. Compositional mapping

2.1 Compositional mapping based on physical maps
2.2 Chromosomal compositional mapping at a 400-band resolution
2.3 Chromosomal compositional mapping at a 850-band resolution

Chapter 3. Genes, isochromes and bands in human chromosomes 21 and 22

Chapter 4. Replication timing, recombination and transcription of chromosomal bands
<table>
<thead>
<tr>
<th>Chapter 5. Isochores in the interphase nucleus</th>
<th>209</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Distribution of the GC-richest and GC-poorest isochores in the interphase nucleus of human and chicken</td>
<td>209</td>
</tr>
<tr>
<td>5.2 Different compaction of the human GC-richest and GC-poorest chromosomal regions in interphase nuclei</td>
<td>209</td>
</tr>
<tr>
<td>5.3 The spatial distribution of genes in interphase nuclei</td>
<td>213</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PART 8: THE ORGANIZATION OF PLANT GENOMES</th>
<th>217</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 1. The organization of the nuclear genome of plants</td>
<td>219</td>
</tr>
<tr>
<td>Chapter 2. Two classes of genes in plants</td>
<td>225</td>
</tr>
<tr>
<td>Chapter 3. Gene distribution in the genomes of plants</td>
<td>227</td>
</tr>
<tr>
<td>3.1 The gene space in the genomes of <em>Gramineae</em></td>
<td>227</td>
</tr>
<tr>
<td>3.2 Misunderstandings about the gene space of <em>Gramineae</em></td>
<td>231</td>
</tr>
<tr>
<td>3.3 The gene space of other plants</td>
<td>233</td>
</tr>
<tr>
<td>3.4 Distribution of genes in the genome of <em>Arabidopsis</em></td>
<td>234</td>
</tr>
<tr>
<td>3.5 A comparison of the genomes of <em>Arabidopsis</em> and <em>Gramineae</em></td>
<td>236</td>
</tr>
<tr>
<td>3.6 The bimodal gene distribution in the tobacco genome</td>
<td>239</td>
</tr>
<tr>
<td>3.7 Methylation patterns in the nuclear genomes of plants</td>
<td>239</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PART 9: THE COMPOSITIONAL PATTERNS OF THE GENOMES OF INVERTEBRATES, UNICELLULAR EUKARYOTES AND PROKARYOTES</th>
<th>241</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 1. The genome of a Urochordate, <em>Ciona intestinalis</em></td>
<td>243</td>
</tr>
<tr>
<td>Chapter 2. The genome of <em>Drosophila melanogaster</em></td>
<td>247</td>
</tr>
<tr>
<td>Chapter 3. The genome of <em>Caenorhabditis elegans</em></td>
<td>251</td>
</tr>
<tr>
<td>Chapter 4. The nuclear genome of unicellular eukaryotes</td>
<td>253</td>
</tr>
<tr>
<td>Chapter 5. Compositional heterogeneity in prokaryotic genomes</td>
<td>257</td>
</tr>
<tr>
<td>5.1 CsCl gradient ultracentrifugation and traditional fixed-length window analysis</td>
<td>257</td>
</tr>
<tr>
<td>5.2 Generalized fixed-length window approaches</td>
<td>257</td>
</tr>
<tr>
<td>5.3 Intrinsic segmentation methods</td>
<td>259</td>
</tr>
<tr>
<td>5.4 Does intragenomic heterogeneity in <em>E. coli</em> arise from exogenous or endogenous DNA?</td>
<td>262</td>
</tr>
<tr>
<td>5.5 Inter- and intra-genomic GC distributions</td>
<td>263</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PART 10: GENE COMPOSITION AND PROTEIN STRUCTURE</th>
<th>265</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 1. The universal correlations</td>
<td>267</td>
</tr>
<tr>
<td>Chapter 2. The universal correlations and the hydrophobicity of proteins</td>
<td>271</td>
</tr>
<tr>
<td>Chapter 3. The universal correlation and imaginary genes</td>
<td>279</td>
</tr>
<tr>
<td>Chapter 4. Compositional gene landscapes</td>
<td>281</td>
</tr>
<tr>
<td>4.1 Large-scale-features of the human gene landscape</td>
<td>281</td>
</tr>
<tr>
<td>4.2 Gene landscapes correspond to protein landscapes</td>
<td>283</td>
</tr>
<tr>
<td>4.3 Gene landscapes correspond to experimentally determined DNA landscapes</td>
<td>283</td>
</tr>
<tr>
<td>Chapter 5. Nucleotide substitutions and composition in coding sequences. Correlations with protein structure</td>
<td>285</td>
</tr>
<tr>
<td>5.1 Synonymous and nonsynonymous substitution rates in mammalian genes are correlated with each other</td>
<td>285</td>
</tr>
</tbody>
</table>
5.2 Synonymous and nonsynonymous substitution rates are correlated with protein structure ................................................................. 287
5.3 Synonymous and nonsynonymous substitution rates are correlated with protein structure: an intragenic analysis of the Leishmania GP63 genes .... 287
5.4 Base compositions at nonsynonymous positions are correlated with protein structure and with the genetic code .................................................. 288
5.5 Base composition at synonymous positions are correlated with protein structure .......................................................... 291

PART 11: THE COMPOSITIONAL EVOLUTION OF VERTEBRATE GENOMES ..... 293
Chapter 1. Two modes of evolution in vertebrates .................................................. 295
Chapter 2. The maintenance of compositional patterns ........................................... 297
  2.1 The maintenance of the compositional patterns of warm-blooded vertebrates 297
  2.2 The conservative mode of evolution and codon usage .................................. 298
  2.3 Mutational biases in the human genome .................................................... 300
Chapter 3. The two major compositional shifts in vertebrate genomes ................... 303
  3.1 The major shifts ....................................................................................... 303
  3.2 Compositional constraints and codon usage ................................................ 310
  3.3 Other changes accompanying the major shifts ........................................... 313
Chapter 4. The minor shift of murids ................................................................. 317
  4.1 Differences in the compositional patterns of murids and other mammals .... 317
  4.2 Isochore conservation in the MHC loci of human and mouse ...................... 318
  4.3 The increased mutational input in murids .................................................. 322
Chapter 5. The whole-genome shifts of vertebrates ............................................. 323

PART 12: NATURAL SELECTION AND GENETIC DRIFT IN GENOME EVOLUTION: THE NEO-SELECTIONIST MODEL ........................................ 325
Chapter 1. Molecular evolution theories and vertebrate genomics ....................... 327
  1.1 Molecular evolution theories .................................................................. 327
  1.2 Structural genomics of vertebrates ......................................................... 329
  1.3 Our previous conclusions ....................................................................... 331
Chapter 2. Natural selection in the maintenance of compositional patterns of vertebrate genomes: the neo-selectionist model ...................... 333
Chapter 3. Natural selection in the major shifts ................................................... 337
Chapter 4. The causes of the major shifts ........................................................ 339
  4.1 Compositional changes and natural selection ............................................ 339
  4.2 The thermodynamic stability hypothesis: DNA results .............................. 339
  4.3 The thermodynamic stability hypothesis: RNA results ............................. 347
  4.4 The thermodynamic stability hypothesis: Protein results ......................... 347
  4.5 The *primum movens* problem .................................................................. 351
Chapter 5. Objections to selection ................................................................. 353
Chapter 6. Alternative explanations for the major shifts .................................... 361
Chapter 7. Natural selection and the “whole genome” shifts of prokaryotes and eukaryotes 367

RECAPITULATION .................................................................................. 369
1. Structural genomics of warm-blooded vertebrates ........................................ 370
2. Chromosomes and interphase nuclei ............................................................ 374
3. Comparative and evolutionary genomics of vertebrates ................................ 375
4. The eukaryotic genome ............................................................................. 382
5. The prokaryotic genome ................................................................. 383

CONCLUSIONS .................................................................................. 385

Abbreviations ................................................................................... 389
References ......................................................................................... 391

SUBJECT INDEX .................................................................................. 435
AN UPDATE ......................................................................................... 440

* Except where indicated otherwise, all figure legends are verbatim transcriptions of the original ones.
Part 1
Introduction
**Introduction**

1.1. The genome: a short history of different views

Since our starting point was the analysis of the organization of the eukaryotic genome, it may be appropriate to begin this introduction with a brief history of the different views concerning the genome. The term genome was coined over eighty years ago by Hans Winkler (1920), a Professor of Botany at the University of Hamburg, to designate the **haploid chromosome set**. Interestingly, the term genome was associated with eukaryotes from the beginning. The definition of Winkler was, however, a purely operational definition. This was in contrast to the older definition of gene (Johannsen, 1909), which was a conceptual definition. Indeed, the gene was defined as a unit of the genetic material localized in the chromosomes, and was originally supposed to be at the same time the ultimate unit of inheritance, of phenotypic difference and of mutation.

The term genome was not as successful as the term gene and, in fact, was forgotten for many years. Its utility became evident almost thirty years later, when Boivin et al. (1948) and Vendrely and Vendrely (1948) discovered that the amount of DNA per cell was a characteristic, constant feature of a given species and that somatic cells had a double amount of DNA compared to germ cells, two points confirmed and expanded later (Mitsky and Ris, 1949, 1951). The amount of DNA in haploid cells from organisms belonging to the same species was called **c-value** (Swift, 1950) for constant value, or **genome size** (Hinegardner, 1976; see Cavalier-Smith, 1985, and Petrov, 2001, for reviews). The identical functional potential of the genomes from all cells of a eukaryotic organism was demonstrated later by Gurdon (1962).

Between the end of the 1940's and the end of the 1960's, when the **prokaryotic paradigm** suggested that the eukaryotic genome was essentially made of genes, the word genome was considered to indicate the **sum total of genes**. In fact, the belief was widespread that, taking into account the different size of bacterial and human genomes, the human genome comprised one million genes.

The **large variability of genome sizes**, even among phylogenetically close species, and the discovery of **repeated sequences** led to the idea that genes only represented a part, and often a very small part, of the eukaryotic genome (see **Table 1.1**). The meaning of the word genome changed once more to indicate the **sum total of coding and non-coding sequences**. At this point, at the end of the 1960's, the term genome started its real career. Its increasing popularity accompanied the development of genome projects which began in the 1980's.

A crucial question is whether the eukaryotic genome is fully described by Winkler's definition (as proposed, often only implicitly, in all current textbooks of Molecular or Cell Biology, Genetics, Evolution), and by its subsequent modifications, or whether it is more than the sum of its parts. This dilemma may also be phrased differently, whether the component parts of the genome are endowed with simple **additive properties**, or with **cooperative properties**.

The first view, in which genes were visualized as distributed at random in the bulk of non-coding DNA, which would be "**junk DNA**" (Ohno, 1972) or, at least, "**selfish DNA**"
TABLE 1.1
Genome size, coding sequences and gene numbers in some representative organisms.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome size(^a) Mb(^b)</th>
<th>Coding sequences %</th>
<th>Gene number(^a)</th>
<th>kb/gene(^c) (^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haemophilus</td>
<td>2</td>
<td>85</td>
<td>2,000</td>
<td>1</td>
</tr>
<tr>
<td>Yeast</td>
<td>12</td>
<td>70</td>
<td>6,000</td>
<td>2</td>
</tr>
<tr>
<td>Human</td>
<td>3,200</td>
<td>2</td>
<td>32,000</td>
<td>100</td>
</tr>
</tbody>
</table>

\(^a\) in approximate figures  
\(^b\) kb, kilobases, or thousands of base pairs, bp; Mb, megabases, or millions of bp.

(Doolittle and Sapienza, 1980; Orgel and Crick, 1980), could be paraphrased from Mayr (1976) as the “bean-bag view” of the genome.

The second view of the eukaryotic genome as an integrated ensemble, defended here, is based on the notion that the genome is more than the sum of its parts, because structural, functional and evolutionary interactions occur among different regions of the genome and, more specifically, between coding and non-coding sequences.

To summarize, we have witnessed several different views of the eukaryotic genome: the purely operational view of Winkler, the prokaryotic paradigm, the “bean-bag” view and, finally, the integrated ensemble view. This latter view could, however, only be justified if one could define properties that are specific for the genome as a whole. The main achievements of our work were that we could define such genome properties and that we were able to build a coherent and comprehensive picture, which essentially emerged from an approach jointly based on molecular genetics and molecular evolution. Our main discoveries, concerning the compositional compartmentalization of the vertebrate genome into a mosaic of isochores, the genome phenotypes, the genomic code, the bimodal distribution of genes and its correlation with functional properties could not be accounted for by the classical selection theory or by the mutation-random drift theory, Kimura’s neutral theory of evolution. This led us to investigate further the roles of natural selection and random drift in genome evolution, to propose a paradigm shift which could reconcile the neutral theory with our view of the dominant role played by natural selection in genome evolution and to formulate a neo-selectionist model.

1.2. Population genetics and molecular evolution

It has been stated (Li and Graur, 1991; Li, 1997; Graur and Li, 2000) that molecular evolution “has its roots in two disparate disciplines: population genetics and molecular biology. Population genetics provides the theoretical foundation for the study of the evolutionary process, while molecular biology provides the empirical data”. In my opinion, this concept should be modified. First, population genetics has a number of intrinsic limitations that are best illustrated by its incapability to solve the neutralist-selectionist debate (see, for example, Hey, 1999; Kondrashov, 2000), to quote one example which is at the
centre of the subject matter of this book. Second, molecular biology has played a role that is much more important than providing empirical data, which is rather the task of mapping, sequencing, etc. In fact, molecular biology, which arose in the middle of the past century from the disciplines of biochemistry and molecular structure, has revolutionized biology, one field after the other. One may wonder where we would be in genetics, immunology, virology, cell biology, if molecular biology had not invaded and pervaded those disciplines. Interestingly, starting with the epochal paper by Zuckerkandl and Pauling (1962), evolutionary studies are those that have undergone the deepest changes as a consequence of the development of molecular genetics, as strongly stressed by Kimura (1983), a population geneticist. Indeed, evolutionary genomics, which applies the molecular biology approach to the study of genome evolution, is progressively transforming the most speculative field of biology into the most rigorous one. This book shows how the structural genomics results obtained in our laboratory led not only to a better understanding of genome organization, but also to advances in evolutionary genomics which, in turn, opened the way to the solution of the neutralist-selectionist debate. Incidentally, it is because of the overwhelming role played by the molecular approach in our work that this book is published in a series called *New Comprehensive Biochemistry*.

### 1.3. Three remarks on terminology

I will not use in this book the expression **GC content**, preferring **GC level** or, simply, **GC** (which is defined as the molar ratio of guanosine and cytidine in DNA; see *Abbreviations*). Indeed, one can talk about a content only if there is a container. In the case of DNA, the nucleotides are not contained in DNA, they form DNA.

The counterpart of AT is not CG (as used by some authors) but GC, because what matters here is not the alphabetical order but the purine-pyrimidine order.

Finally, let me stress that by structural genomics I mean what used to be called nucleotide sequence organization or genome organization and not, as recently suggested (see, for example, Stevens et al., 2001; Baker and Sali, 2001), protein structure (although the latter also enters into the picture).

### 1.4. A brief chronology of our investigations

I will present here a short narrative of our research, concentrating on its early phases, because they are not dealt with in the book. My research career started in 1951, when I rang the bell of the Medicinska Nobel Institute, Department of Biochemistry in Stockholm, and was accepted by Professor Hugo Theorell as a summer student to start my thesis in biochemistry in view of a Medical Degree at the University of Padova. After the defence of my thesis, I spent two years in the Italian Air Force, during which time I kept in touch with the Department of Biochemistry of the University of Padova and started studying Physics (I later obtained a “Libera Docenza” in Physical Biochemistry in 1962 and a “Doctorat d’Etat ès-Sciences Physiques” in 1967 with Jacques Monod as the Chairman of the Jury). I then moved to the Biochemistry Department of the University of Pavia. My
work there on the physico-chemical properties of mucopolysaccharides led me to visit the Centre de Recherche sur les Macromolécules of Strasbourg, directed by Professor Charles Sadron. As a consequence of this brief visit, in 1956 I left Italy to work in Strasbourg for several months. After six more months, during which I pursued investigations on mucopolysaccharides with Professor Frank Happey in Bradford, I joined, as a post-doctoral fellow, the National Research Council of Canada in Ottawa, where I worked on lipoproteins with Dr. W.H. Cook between 1957 and 1959. In those years, the return journey by boat from New York to Genoa took 12 days, a time long enough to think about the future and to make the decision to devote myself to the study of DNA. At that time, this was the research subject of only a handful of laboratories in the world (the Watson and Crick paper of 1953 was barely quoted in the 1950’s). I knew that this was going to be a vast enterprise, but I could not imagine that my work in the field was going to span more than 40 years. It was, however, the best decision, because it allowed me to take an active role in an adventure that led us from the double helix to the human genome sequence through the golden era of molecular biology.

When I started working in the pleasant environment of the Centre de Recherche sur les Macromolécules in Strasbourg in 1959, it was clear to me that two tools were needed to understand the way the genome of eukaryotes was organized: enzymes that were able to cut DNA into large fragments and fractionation methods that were able to separate the fragments. I therefore embarked, at the age of 30, on two research projects along those lines.

As an enzyme, I chose acid DNase (which we had isolated for the first time), because preliminary experiments indicated that this enzyme led to the degradation of DNA into very large fragments of about 1 kb (Bernardi et al., 1960). The study of this and other DNases provided the first demonstration that these enzymes recognize short DNA sequences, contrary to the prevailing view (Laskowski, 1971, 1982) that they lacked specificity. It also indicated that acid DNase could cut both strands of native DNA at the same time.

![Diagram of DNA fragmentation](image-url)

Figure 1.1. Analysis of termini: a number of sequences, shown as tetranucleotides and numbered 1 to 5, are recognized and split with different $K_m$ (and/or $V_{max}$), indicated by $K_1$, $K_2$, etc. Terminal and penultimate nucleotides were isolated from the resulting oligonucleotides, and their base compositions (see Fig. 1.2) were determined. (From Bernardi et al., 1973).
Figure 1.2. Scheme of a tetranucleotide split by a DNase at the position indicated by an arrow. Average nucleotide composition at the two terminal positions, W and Z, and at the two penultimate positions, X and Y, were determined using methods that we developed for their isolation and analysis. (From Bernardi et al., 1973).

In turn, this suggested that symmetrical sequences were recognized and split, so implying a symmetry in the DNase molecule, which was, in fact, an allosteric dimer (Bernardi 1965a). Even if these early studies (summarized in Bernardi, 1971) may need some revision, in many respects acid DNase was a prefiguration of the (class II) restriction enzymes which were discovered ten years later (Smith and Wilcox, 1970), and which made it obsolete for the purpose of cutting DNA into large pieces, because of their strict sequence specificity. One point which acid DNase, with its lower yet defined specificity, allowed us to assess was, however, the frequency of the sequences that the enzyme could recognize and split, namely the average composition of the terminal and penultimate nucleotides (the termini) on each side of the cuts (Figs. 1.1 and 1.2).

These nucleotides had compositions that were characteristic of the DNA under study (and of the DNase used). Since the percentages of the termini formed by DNases from bacterial DNAs were linearly related to their GC levels (Fig. 1.3A), a useful way to show the results from the DNAs under examination was to plot difference histograms like those of Fig. 1.3B. This approach extended the nearest neighbour analysis of dinucleotide frequencies (Josse et al., 1961) to a frequency approach involving the sequences, at least four

---

**Figure 1.3.** A. The percentages of the four nucleotides, A (circles), G (squares), C (diamonds) and T (triangles) in the 3'-terminal, 5'-terminal and 5'-penultimate nucleotides formed by the spleen and the snail DNase from bacterial DNAs (*Haemophilus influenzae*, 38% GC; *Escherichia coli*, 51% GC; *Micrococcus luteus* 72% GC) are plotted against the GC level of DNAs. Values obtained at an average chain length of 15 nucleotides were used. B. Deviation patterns of three repetitive DNAs. The histograms show the differences between the composition of termini formed from guinea pig satellite, mouse satellite and yeast mitochondrial DNAs by spleen and snail DNase and the compositions expected for bacterial DNAs having the same GC level; tl, terminal; pt, penultimate. (From Bernardi et al., 1973).
nucleotides long, that had been recognized and split by the enzyme (Bernardi et al., 1973). This allowed us to see specific patterns in different DNAs, not only in repetitive DNAs (see Fig 1.3B), but also in human “main-band” DNA and its major components (see Fig. 3.8). We later applied frequency methods to oligonucleotides from the mitochondrial genome of yeast (see Part 2). Interestingly, frequency methods involving di- to tetra-nucleotides have now been revived using self-assembly approaches and complete genomic sequences (Abe et al., 2003).

As a fractionation method, I developed chromatography of nucleic acids on hydroxyapatite, a calcium phosphate which had been used by Tiselius et al. (1956) for the fractionation of proteins. Previous observations (Bernardi and Cook, 1960a,b,c) that hydroxyapatite was particularly good as a chromatographic substrate for fractionating of phospholipoproteins characterized by different phosphorylation levels convinced me to try it on DNA. The main discovery was that hydroxyapatite could fractionate single- from double-stranded DNA (Fig. 1.4 left panel), the former being eluted by a lower phosphate

![Figure 1.4.](image_url)

**Figure 1.4.** **Left panel:** Gradient elution by phosphate buffer in the presence of 1 per cent formaldehyde of bovine DNA: A, native DNA; B, heat-denatured (100°C) DNA; C, a 1:1 mixture of native and heat-denatured (100°C) DNA. (From Bernardi, 1965b). **Right panel:** Chromatography of DNA preparations A from wild-type yeast cells; B from a cytoplasmic petite mutant. The three peaks eluted by the phosphate gradient correspond to RNA, nuclear DNA (a) and mitochondrial DNA (b). W and G indicate washing and gradient. (From Bernardi et al., 1968)
buffer molarity compared to the latter (Bernardi, 1962, 1965b). The reason for this separation was that the random coil form of denatured DNA had fewer phosphate groups available for the binding to calcium sites of hydroxyapatite compared to the extended, relatively rigid, double-stranded, native DNA. This also explained the separations of compact, supercoiled from relaxed polyoma virus DNA (Bourgaux and Bourgaux-Ramoisy, 1967), of RNA from DNA (see for example Fig. 1.4, right panel), and, also, of denatured from native proteins (Bernardi et al., 1972a). The fractionation of single- and double-stranded DNA on hydroxyapatite provided an extremely powerful tool to study the kinetics of reassociation of denatured DNA and to demonstrate the existence of repeated sequences in eukaryotic DNAs (Britten and Kohne, 1968).

Not taking into consideration the case in which secondary (single-stranded DNA from \( \Phi X \) 174 phage) or tertiary structures (twisted, circular DNA from polyoma virus) are grossly different and cause a different chromatographic behaviour, it is evident that differences in nucleotide sequences in otherwise similar DNAs (the similarity concerning the double-strandedness, the molecular weight, the linear or open circular configuration, and also the nucleotide composition) may be sufficient to determine different elution molarities. Indeed, DNAs containing short repetitive sequences may in general show particular elution molarities. For example, DNAs comprising alternating dAT:dAT and non-alternating dA:dT structures, like mitochondrial DNAs from *Saccharomyces cerevisiae* (Bernardi et al., 1968, 1970; see Fig. 1.4 right panel, and Part 2), *Euglena gracilis* (Stutz and Bernardi, 1972; Fonty et al., 1975), *Ustilago cynodontis* (Mery-Drugeon et al., 1981) and chloroplast DNA from *Euglena gracilis* (Schmitt et al., 1981; Heizmann et al., 1981) showed high elution molarities compared to nuclear DNAs. In spite of such separations, in general the resolving power of hydroxyapatite was not good enough for fine fractionations of high molecular weight DNA according to base composition. The use of hydroxyapatite was, therefore, limited (not without problems) to reassociation kinetics experiments (see Table 3.1). Hydroxyapatite chromatography was, however, not abandoned. A long series of theoretical investigations started in our laboratory were carried out by Tsutomu Kawasaki, who summarized them in a book (Kawasaki, 2003).

In 1966 I turned, therefore, to equilibrium centrifugation of DNA in density gradients, in the presence of sequence-specific DNA ligands, as a fractionation approach based on the frequency of the oligonucleotides that could bind the ligand. This worked beautifully not only for separating satellite DNAs (Corneo et al., 1968), but also, to our surprise, for fractionating main-band bovine DNA into three major DNA components (in addition to several satellite and minor DNA components; Filipski et al., 1973). This work was done in Paris where I had moved from Strasbourg in 1970 to head the Laboratoire de Génétique Moléculaire at the new Institut de Recherches en Biologie Moléculaire (later called, at my suggestion, Institut Jacques Monod, after the name of its founder).

The finding of a striking, discontinuous, compositional heterogeneity at the macromolecular level in the bovine genome led us to investigate other eukaryotic DNAs (Thiery et al., 1976) and to show (Macaya et al., 1976) that the genome of vertebrates was a mosaic of isochores (for compositionally equal landscapes), namely of chromosomal DNA regions originally estimated as over 300 kb long, on the average, which were fairly homogeneous in composition, and belonged to different families characterized by different average GC levels. In 1984, we presented at the FEBS Meeting in Moscow the crucial discovery that
the vertebrate genome was characterized by compositional correlations linking coding and non-coding sequences (Bernardi et al., 1985a,b). Indeed, this changed the previous definitions of the genome (see Section 1.1) into that of a system whose properties were not simply the sum of the properties of its constitutive parts.

In 1966, I also started a long-term project on the molecular genetics of yeast mitochondria. This model system provided important results as far as both the organization and the evolution of eukaryotic genomes are concerned. These findings will be briefly presented and discussed in Part 2 of this book. In the 1960's we also performed a series of investigations on the physical chemistry of DNA (see, for instance, Froelich et al., 1963; Freund and Bernardi, 1963) and on transforming DNA from Haemophilus influenzae (Chevallier and Bernardi, 1965, 1968; Bernardi and Bach, 1968; Kopecka et al., 1973).

A subsequent major step in the history of our laboratory was its increasing involvement (starting in 1980) in both evolutionary aspects of genome organization and computer approaches to the study of nucleotide sequences. At the beginning of 1998, I moved to the Stazione Zoologica of Naples, where I started a Laboratory of Molecular Evolution, fulfilling, more than a century later, the dream of Anton Dohrn, my genial predecessor, who founded the Stazione in 1872 in order to carry out research with the aim of proving that Darwin was right.

1.5. Molecular approaches to the study of the genome

Among the molecular approaches used in our work on the genome organization, equilibrium centrifugation of DNA in density gradients played a key role and allowed us to analyse and fractionate DNA and to characterize the fractions so obtained. It is, therefore, useful to briefly illustrate it. Another good reason to devote some space to this approach is that it is not widely used and not well understood, as shown, for example, by the following statements: (i) “Physical methods have suggested that G+C compositional heterogeneity is less marked in poikilothermic animals (Bernardi, 2000), and this is confirmed by our large-scale sequence analysis” (Aparicio et al., 2002). (ii) “Studies of the compositional patterns of the completely sequenced human genome (International Human Genome Sequencing Consortium, 2001) confirmed earlier indirect results (Bernardi et al., 1985) suggesting that an average GC content in both coding and non-coding DNA varies along the chromosome in a non-random way” (Filipski and Mucha, 2002; our underlinings). Indeed, CsCl centrifugation, although it might be considered an indirect method, did not simply suggest, but actually demonstrated the points under consideration. Interestingly, the compositional heterogeneity of the vertebrate genome was quantified more than a quarter of a century before its “confirmation” by sequencing (Filipski et al., 1973, Thiery et al., 1976, Macaya et al., 1976; see Clay et al., 2003a for additional information).

Fig. 1.5 presents a simple scheme concerning the information that can be obtained by centrifuging DNA to equilibrium in an analytical CsCl density gradient. Scanning the DNA band formed in the centrifuge cell produces an absorbancy profile (at 260 nm) which allows estimating of the modal buoyant density $\rho_0$ (the density at the peak), the mean (average) buoyant density $<\rho>$, the asymmetry of the profile, $<\rho>-\rho_0$, and the heterogeneity, H.
More precisely (see Thiery et al., 1976), in order to calculate the buoyant density, \( \rho \), at any point of abscissa \( r \) from the rotation axis, the relationship

\[
\rho = \rho_\kappa - \frac{\omega^2}{2\beta_0} (r^2 - r' \rho_\kappa^2 - r^2)
\]  

is used, where the subscript \( \kappa \) refers to the marker, \( \omega \) is the angular velocity in radians \( s^{-1} \), and \( \beta_0 \) was taken as equal to 1.19 \( \times \) 109 cm\(^5\) g s\(^{-1}\) (Ifft et al., 1961). Under such conditions, using phage 2C DNA (\( \rho = 1.742 \) g/cm\(^3\); Szybalski, 1968) as a density marker, a reproducible modal buoyant density, \( \rho_0 \) (density at the peak maximum, located at a distance \( \rho_0 \) from the rotation axis), of 1.7103 g/cm\(^3\) is obtained for \( E. coli \) DNA.

The mean buoyant density, \( <\rho> \), is calculated from the first moment of the band profile about the center of rotation:

\[
< r > = \frac{\int_0^\infty r \, c \, d \, r}{\int_0^\infty c \, d \, r}
\]

and from eqn (1), \( c \) being the DNA concentration at point of abscissa \( r \).

The variance of the profile, \( <\delta^2> \), is equal to the second moment about the mean

\[
< \delta^2 > = m_2 = \frac{\int_{cd\delta} \delta \, d\delta}{\int_{cd\delta} d\delta}
\]

with \( \delta = r - < r > \).

Integrals of eqns (2) and (3) can be calculated using Simpson's rule, with radial steps corresponding to 0.04 or 0.08 mm in cell dimensions.

The asymmetry, \( A \), of CsCl main bands was determined as \( A = <\rho> - \rho_0 \).

The intermolecular compositional heterogeneity, \( H \), of CsCl main bands can be calculated according to Schmid and Hearst (1972).
\[ H = \left[ \frac{G \rho_0^2 \omega^4 \delta^2}{RT \rho_0} + \frac{1}{M_3} - \frac{RT \rho_0}{\beta_0^2 a^2 G} \right]^{1/2} \]  

where \( G \) is the buoyancy factor, \( M_3 \) is the molecular weight of the dry Cs salt of DNA, \( R \) is the universal gas constant, \( T \) is the absolute temperature, and \( a \) is the slope, 0.098, of the relationship of Schildkraut et al. (1962) between buoyant density and GC,

\[ \rho = 1.66 + a\%GC \]

It should be stressed that, while in vertebrate DNAs there is a significant correlation between asymmetry and heterogeneity (see Fig. 4.2D), this is not necessarily always the case. Indeed, while an asymmetric profile is always heterogeneous, a heterogeneous profile is not necessarily asymmetric (as when an asymmetry on the GC-rich side is compensated by an asymmetry on the GC-poor side of the profile).

Sedimentation velocities of DNAs can be determined as described by Prunell and Bernardi (1973), and molecular weights then calculated from the sedimentation coefficients \( s_{20,w} \), using the relationship of Eigner and Doty (1965)

\[ s_{20,w} = 0.034M^{0.405} \]

Molecular weight of DNA is currently expressed in bp (base pairs) and its multiples kb, Mb, Gb (see legend of Fig. 1.1). Older data are often given in daltons, 1kb being about equal to \( 6.18 \cdot 10^5 \) daltons. Genome size is often given in picograms (1 pg = \( 10^{12} \)g; 1 pg \( \approx 0.98 \cdot 10^3 \)kb \( \approx 6.02 \cdot 10^{14} \) daltons).

When centrifuged to equilibrium in analytical CsCl density gradients, undegraded viral genomes show unimodal, narrow symmetrical bands, as expected for a population of compositionally identical DNA molecules (see Fig. 1.6 for an example).

In the case of the much larger bacterial genomes (~ 2–5 Mb), DNA preparations consist
of fragments of different sizes. At average sizes above 50 kb, DNA fragments from *E. coli* DNA are essentially identical in composition (Yamagishi, 1970), hence the conclusion that bacterial genomes are characterized by a fairly homogeneous composition above a certain fragment size. This conclusion (to be discussed in detail in Part 9) was reinforced by the comparison with the then “standard” eukaryotic DNA, calf thymus DNA, which showed a broad, asymmetrical band (Meselson et al., 1957; Sueoka, 1959). In fact, the very strong asymmetry of the band was mostly due to the presence (see Part 3) of eight GC-rich satellite DNAs, six or seven (according to the resolution) of which were not separated from the “main-band” of bovine DNA (see Figs. 3.5 and 3.6). Satellite DNAs are components of eukaryotic DNAs (later shown to be made up of tandem oligonucleotides) that appear as separate bands accompanying the main band, or as shoulders of the main band (Kit, 1961; Sueoka, 1961), although some satellite DNAs remain cryptic, unseparated from the main band (see below).

When we started to look into the problem of fractionating eukaryotic DNAs, the first mammalian satellite DNA had just been isolated by preparative CsCl density gradient centrifugation from the mouse genome (Waring and Britten, 1966; Bond et al., 1967). The separation of mouse satellite (*ρ* = 1.690 g/cm³) from main-band DNA (*ρ* = 1.700 g/cm³) was not an easy one, because the two buoyant densities in CsCl were rather close. The guinea pig satellite was an even more difficult case because its buoyant density was 1.705 g/cm³ vs. 1.700 g/cm³ for the main band. We decided, therefore, to attempt the isolation of both mouse and guinea pig satellite DNAs by using centrifugation in Cs₂SO₄/Ag⁺ density gradient, an approach just developed by Jensen and Davidson (1966) to separate the “AT satellite” from crab DNA. This approach was successful. Indeed, in such gradients the mouse satellite was well separated from the main band (by 45 mg/cm³ instead of 10 mg/cm³ in CsCl; see Fig. 1.7 and Table 1.2). In the case of guinea pig, a major, heavy (GC-
TABLE 1.2
Modal buoyant densities and GC levels of main-band and satellite DNAs from mouse and guinea pig\(^a\)

<table>
<thead>
<tr>
<th>Bands</th>
<th>Mouse DNA</th>
<th>Guinea Pig DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Main</td>
<td>Satellite</td>
</tr>
<tr>
<td>CsCl</td>
<td>1.701</td>
<td>1.690</td>
</tr>
<tr>
<td>Cs(_2)SO(_4)/Ag(^+)</td>
<td>1.501</td>
<td>1.456</td>
</tr>
<tr>
<td>CsCl buoyant density(^b)</td>
<td>40.0</td>
<td>30.6</td>
</tr>
<tr>
<td>Melting temperature(^c)</td>
<td>40.0</td>
<td>40.8</td>
</tr>
<tr>
<td>Nucleoside analysis</td>
<td>40.0</td>
<td>35.2</td>
</tr>
</tbody>
</table>

\(^a\) From Corneo et al. (1968)
\(^b\) Calculated according to Schildkraut et al. (1962)
\(^c\) Calculated according to Marmur and Doty (1962)

rich), and two minor, light (GC-poor) satellite bands showed up (Corneo et al., 1968; see Fig. 1.7). While the former appeared as a shoulder on the CsCl profile, the latter were “cryptic” satellite DNAs that could not be detected in CsCl. As in the case of the mouse satellite, the major guinea pig satellite was better separated from main-band DNA in Cs\(_2\)SO\(_4\)/Ag\(^+\) (by 34 mg/cm\(^3\) instead of only 6 mg/cm\(^3\) in CsCl).

The mouse satellite and the guinea pig major satellite were then isolated by preparative ultracentrifugation in Cs\(_2\)SO\(_4\)/Ag\(^+\) and their strands were separated by ultracentrifugation in alkaline CsCl. The satellite DNAs and their separated strands were enzymatically degraded to nucleosides (thus avoiding the losses associated with the chemical hydrolysis of DNA to bases) using, in succession, a series of enzymes isolated in our laboratory, spleen acid DNAse (Bernardi et al., 1966), spleen acid exonuclease (Bernardi and Bernardi, 1968) and spleen acid phosphomonoesterase II (Chersi et al., 1966). Nucleosides were then separated on Bio-Gel P-2 columns and quantified (Carrara and Bernardi, 1968; Piperno and Bernardi, 1971). Our analyses revealed that the two satellites from mouse and guinea pig only differed by 3% GC, instead of the 15% predicted by the relationship of Schildkraut et al. (1962). Moreover, the satellites were also characterized by an asymmetry of base compositions on the two strands. This was true not only for mouse satellite (as already seen by Flamm et al., 1967), but also, and much more so, for the major guinea pig satellite, which exhibited a “light strand” with less than 3% G. Finally, buoyant density in CsCl underestimated the GC level in mouse satellite, but overestimated it in guinea pig satellite (see Table 1.2). This was a clear indication that the basis for the separation was the differential binding of silver ions to the short sequences that made up the satellite DNAs under consideration (see Fig. 1.8). In fact, the high resolution depended upon such differential binding, since, per se, the resolving power of Cs\(_2\)SO\(_4\) density gradients for DNA molecules of different base composition is much lower than that of CsCl gradients (Szybalski, 1968; Schmid and Hearst, 1972).
A novel observation was, however, that the base composition from the nucleoside analysis was very different from that calculated from CsCl buoyant density (Schildkraut et al., 1962), or from melting temperature (Marmur and Doty, 1962). Interestingly, while this was true for satellite DNAs, it was not for main band DNA (see Table 1.2). Having ruled out the presence of rare bases as the cause for the discrepancies, the explanation which was proposed was that “satellite DNAs are conformationally slightly different from main-band DNAs, the differences being related to their peculiar nucleotide sequences” (Corneo et al., 1968). Indeed, the sequence dependence of buoyant density in CsCl is most probably due to a different water binding by different short sequences in the DNA molecules (the bases themselves having identical “dry” densities). In connection with this explanation, it may be recalled that non-alternating poly (dA) (dT), alternating poly (dAT) (dAT) and poly (dG) (dC) were already known to show “anomalous” buoyant densities (Schildkraut et al., 1962), melting temperature (Marmur and Doty, 1962) and optical rotatory dispersion (Samejima and Yang, 1965). Indeed, their properties were different from those of the prokaryotic DNAs used to establish the relationships between base composition and physical properties. These anomalies were, generally, attributed to the extreme compositions of the synthetic polynucleotides, although Wells and Blair (1967) had shown a sequence-dependent behaviour of synthetic polynucleotides made up of repeated trinucleotides. Similar anomalous physical properties were also found (Bernardi et al., 1968, 1970; Bernardi and Timasheff, 1970) for yeast mitochondrial DNAs from wild-type cells (18% GC) and cytoplasmic “petite colonie” mutants (ranging from 4% to 18% GC), due to their abundance in short alternating and non-alternating AT sequences (see Part 2).

The basic methods that we routinely used to study genome organization were (see also Table 1.3):

(i) Analytical equilibrium ultracentrifugation in CsCl density gradients (see Fig. 1.5).

(ii) Analytical and preparative equilibrium centrifugation in Cs2SO4 density in the presence of sequence-specific DNA ligands (see Fig. 1.7): the two ligands that we used were Ag+ ions at two different pH values and BAMD, bis(acetatomercurimethyl)dioxane, an organic mercurial first synthesized at the end of the 19th century (see Bünemann and Dattagupta, 1973). Two interesting features about BAMD are 1) that the compound is not sensitive to contaminating impurities, which is the case for Ag+; and 2) that one can
TABLE 1.3
Methods for studying nucleotide sequences in DNAs

Indirect methods
1. Sequence-dependent properties of double-stranded DNA
   a. Buoyant density in CsCl
   b. Melting temperature
   c. Ligand binding
   d. Hydroxyapatite binding
2. Sequence-dependent properties of single-stranded DNA
   a. Hydroxyapatite binding
   b. Reassociation kinetics

Direct methods
1. Frequency methods
   a. Depurination
   b. Nearest-neighbour analysis
   c. Analysis of termini released by DNases
2. Mapping
3. Sequencing

* From Bernardi et al. (1973).

Figure 1.9. DNA distribution in a shallow CsCl gradient of A a GC-rich YAC (Yeast Artificial Chromosome) and B of a GC-poor YAC. λ and T4 bacteriophage DNAs were used as density markers. The intensities of the hybridization signals (left ordinate) and the buoyant densities (right ordinate) are plotted against the fractions collected from the gradient. (From De Sario et al., 1995).
shift the optimum resolution range by changing the ligand/nucleotide ratio (this is also true for Ag⁺).

(iii) Preparative equilibrium centrifugation in shallow CsCl density gradients (De Sario et al., 1995; see Fig. 1.9); this approach, although not having quite the same resolving power as the DNA-ligand approach just mentioned, had the advantage of being much less laborious and of avoiding DNA losses, because dialysis to remove the ligand was not needed; DNA fractions could be alkali-denatured, loaded on filters, washed out of CsCl and hybridized with appropriate probes.

(iv) Hybridization of specific, labelled probes on fractionated DNA. This approach not
only localizes the sequence of interest in compositional fractions (Fig. 1.10), but also provides information on the average GC level of a DNA region having a size up to twice the average size of the target DNA molecules (see Fig. 1.11). PCR (polynucleotide chain reaction) was also used on DNA fractions in order to localize specific sequences.

(v) Hybridization of compositional DNA fractions on chromosomes (see Part 7) in the presence of an excess of unlabelled $C_o t^{-1}$ DNA; ($C_o t$ is the product of the initial DNA concentration by the reannealing time).

**In conclusion**, the main experimental approach that we followed to investigate genome organization was based on the most elementary properties of the genome, its nucleotide composition, and its oligonucleotide frequencies. This compositional approach, the only one that was possible before DNA sequencing, is still very useful, in particular for screening large numbers of genomes or genome fractions. Moreover, the compositional approach could be, and indeed was, easily transferred from DNA molecules to DNA sequences, when these became available. The usefulness for the compositional approach was three-fold: (i) the compositional heterogeneity of the genome could not only be detected and assessed, but also used for fractionation purposes, so allowing us to study the properties of compositional DNA fractions; (ii) compositional features could be analysed in very large sequences, such as the human genome sequences; (iii) compositional changes resulting from the evolutionary process could be detected and compared.
Part 2
Lessons from a small dispensable genome,
the mitochondrial genome of yeast
The mitochondrial genome of yeast and the petite mutation

1.1. The “petite colonie” mutation

The mitochondrial genome of yeast is of special interest for two major reasons: (i) because, in contrast to its very compact counterpart in animal cells, it comprises abundant non-coding sequences (Bernardi et al., 1970); this situation is not unique to Saccharomyces cerevisiae since it is shared by other fungi and by Euglena gracilis (as shown by investigations presented in Chapter 3), nor to the mitochondrial genome since the chloroplast genome also shows it; in the case of S. cerevisiae, mitochondrial genome units are about five times larger than the units of animal mitochondrial genomes; (ii) because it is dispensable since S. cerevisiae can survive on fermentable carbon sources; mutants (the cytoplasmic “petite colonie” mutants of Ephrussi; see below) having undergone massive alterations in the nucleotide sequences of their mitochondrial genome (Bernardi et al., 1968; Mehrotra and Mahler, 1968) or having lost it altogether can therefore survive (dispensability also applies to the chloroplast genome of E. gracilis; see Heizmann et al., 1981).

In other words, although very small, the mitochondrial genome of yeast presents features that are common to those of the nuclear genome of eukaryotes. One can therefore study, for instance, the organization, the evolutionary origin and the function of non-coding sequences. Hence, our interest. Moreover, since the genome is dispensable, one can investigate genome changes which are usually incompatible with cell life.

In 1948 Boris Ephrussi gave the first account of investigations on the ‘petite colonie’ mutation in S. cerevisiae (Ephrussi, 1949), which were the starting point of extra-chromosomal genetics. It is difficult to describe the initial observation more clearly than in Ephrussi’s own words: “When a culture of baker's yeast, whether diploid or haploid, is plated, each of the cells gives rise in the course of the next few days to a colony. The great majority of these colonies are of very nearly identical size, but one usually finds also a very small number – say 1 or 2% – of distinctly smaller colonies (Fig. 2.1). These facts suggest that the population of cells which was plated was heterogeneous and that it may be possible to purify it by taking cells from either the big or the small colonies only. The results of such a selection show, however, that cells from the big colonies again and again produce the two types of colonies, while the cells from the small colonies give rise to small colonies only” (Ephrussi, 1953). Besides describing the mutation and its irreversibility, this paper also reported a number of other fundamental observations: (i) that acriflavine treatment increased the number of ‘petite’ mutants from 1–2% to 100% (Fig. 2.1); (ii) that the mutants grew slowly because they could not respire, owing to the loss of their ability to synthesize a whole series of respiratory enzymes; and that in anaerobiosis wild-type cells and petite mutants grew at the same slow rate using fermentative pathways; (iii) that crosses of wild-type cells with petite mutants showed a non-Mendelian segregation of the mutation, in that they led either to wild-type progeny, or to both wild-type cells and petite mutants in different proportions.
depending upon the petite used in the cross; the petite mutants entering the cross were called neutral petites in the first case and suppressive petites in the second one (Ephrussi et al., 1955; see Fig. 2.2). The conclusion drawn by Ephrussi was that wild-type cells and petite mutants differed by "the presence in the former and the absence in the latter of cytoplasmic units endowed with genetic continuity and required for the synthesis of certain respiratory enzymes" (Ephrussi, 1953).

![Figure 2.1](image)

**Figure 2.1.** Colonies formed by baker's yeast on a solid medium. **A.** Colonies of a normal yeast, showing one small colony. **B.** Colonies formed by the same yeast grown in the presence of acriflavine prior to plating (From Ephrussi, 1953).

![Figure 2.2](image)

**Figure 2.2.** **A.** The generation of petite mutants from (cytoplasmic) wild-type cells. **B.** The results of crossing wild-type cells by petite mutants.
1.2. The petite mutation is accompanied by gross alterations of mitochondrial DNA

The cytoplasmic units postulated by Ephrussi in 1948 were only identified as mitochondrial genes 20 years later. Indeed the first hard facts about the molecular basis of the petite mutation were published by Bernardi et al. (1968) and by Mehrotra and Mahler (1968). These authors showed that mitochondrial DNAs from two genetically unrelated, acriflavine-induced, petite mutants had a grossly altered base composition (4% GC) compared to DNA from the parent wild-type strain (18% GC). These findings unequivocally established that massive alterations in the nucleotide sequences of the mitochondrial genome may accompany the petite mutation and be responsible for it. These conclusions were quickly confirmed by more detailed investigations (see the following sections), which probed the structure of mitochondrial DNA from both wild-type cells and petite mutants. Since we were interested in understanding the organization of the mitochondrial genome of yeast and the causes of these massive alterations, we decided to solve these problems using a strictly molecular approach, taking advantage of our ability to isolate mitochondrial DNA in large amounts by chromatography on hydroxyapatite (see Fig. 1.4). Wisely, we did not waste time using the classical genetic approach to study the petite mutation, and confined ourselves to spontaneous (versus ethidium-bromide-induced) mutants.

1.3. The AT spacers and the deletion hypothesis

Mitochondrial DNA from wild-type yeast cells was found (Bernardi et al., 1970; see Fig. 2.3) to be extremely heterogeneous in base composition, about half of it melting at a very low temperature and being almost exclusively formed by long stretches of short alternating AT:AT and non-alternating A:T sequences (the existence of which had already been predicted for the first petite genome investigated; Bernardi et al., 1968), and the rest melting over an extremely broad temperature range. Compared with mitochondrial DNA from wild-type cells, DNAs from three spontaneous suppressive petite mutants were shown to have lower GC levels, to lack a number of DNA stretches that melt at high temperature, and to renature very rapidly (Fig. 2.4). At that time, I interpreted those results as indicating that petite mutants had defective mitochondrial genomes, in which large segments of the parental wild-type genomes were deleted. I also suggested that such deletions arose by a mechanism (Campbell, 1969), involving illegitimate, unequal recombination events in the “AT spacers”, which I supposed to contain sequence repetitions because of their extreme base composition. It was obvious that the loss of any known mitochondrial gene products (ribosomal RNAs, tRNAs, the sub-units of enzymes involved in respiration and oxidative phosphorylation) would have a pleiotropic effect and lead to a loss of respiratory functions. Incidentally, this could also happen as a consequence of mutations in nuclear genes that encode some sub-units of mitochondrial enzymes. These “nuclear petite mutants”, that are characterized by a mendelian inheritance, will not be dealt with here.
Figure 2.3. Differential melting curves obtained with poly (dAT : dAT) (solid circles), poly (dA : dT) (open squares), mitochondrial (open circles) and nuclear DNA from wild-type strain B (solid squares). The ordinate, Y, indicates the increment in relative absorbance per degree; $Y_{max}$ of poly (dA:dT) had a value of 0.71. (From Bernardi et al., 1970).

Figure 2.4. Differential melting curves obtained with mitochondrial DNAs. Every curve is an average from 2 to 5 replicate experiments. Strains A and B are two wild-type strains. Strains a1, a2 and b are spontaneous petite strains. See Fig. 2.3 for other indications. (From Bernardi et al., 1970).
1.4. The petite mutation is due to large deletions

Further work (Bernardi et al., 1972b; Prunell and Bernardi, 1977) showed that the AT spacers formed about 50% of the wild-type mitochondrial genome, had a GC content lower than 5%, were indeed repetitive in nucleotide sequence and were, therefore, likely to be endowed with sequence homology over stretches long enough to allow illegitimate site-specific recombination. In 1974, direct evidence was provided for both a deletion mechanism (Bernardi et al., 1975) and an accompanying amplification of the excised genome segment (Locker et al., 1974; Bernardi et al., 1975). Indeed, only a fraction of the restriction fragments of wild-type mitochondrial DNA were present in petite genomes and these were present in multiple copies per genome unit. These findings disposed of a number of strange ad hoc hypotheses put forward to explain the petite mutation and led to the scheme shown in Fig. 2.5 in which the excised segment from the wild-type genome (depicted as circular) becomes the repeat unit of the petite genome. This may, in turn, undergo further deletions leading to secondary petite genomes, characterized by shorter repeat units. Incidentally, the analysis of restriction patterns of mitochondrial DNA from several strains of wild-type cells provided the first unequivocal estimate of the size of the mitochondrial genome unit, about 50 x 10⁶ daltons or about 75 kb (Bernardi et al., 1975), an almost five-fold larger size than that of the mitochondrial genome units of animals, about 17 kb. The strain-related size variation of restriction fragments provided in fact the first example of what was later called “restriction fragment length polymorphism or RFLP”.

Figure 2.5. A. Scheme depicting the excision-amplification process leading to the formation of the genome of a spontaneous petite mutant. A segment of a wild-type mitochondrial genome unit is excised and tandemly amplified into a defective genome unit. This then replicates and segregates into the buds to form the genome of a petite mutant; the petite genome can undergo further excisions leading to the formation of secondary petite genomes. B. Scheme of the excision process leading to the formation of a petite genome unit. Dotted and black areas represent two direct sequence repeats. Excision is followed by a tandem amplification (see A). (From Bernardi, 1983)
Another important advance in our knowledge of the organization of the mitochondrial genome was the discovery of short segments of mitochondrial DNA extremely rich in GC, the "GC clusters" (Bernardi et al., 1976; Prunell et al., 1977; Prunell and Bernardi, 1977). Operationally, two sorts of GC clusters can be distinguished, the (CCGG, GGCC) clusters, present in 60–70 copies per genome unit and recognizable because they are degraded by the restriction enzymes Hpa II and Hae III, and the GC-rich clusters which do not contain CCGG or GGCC sequences, but are often close to (CCGG, GGCC) clusters and to isolated CCGG sequences. A certain number of these clusters were likely to be homologous in sequence. Sequence analyses carried out in several laboratories (see below) indicated that the GC clusters were located in the middle of AT spacers and not at the ends of them, as first suggested as a working hypothesis (Prunell and Bernardi, 1977), so that an overall scheme of the organization of the mitochondrial genome of yeast is that given in Fig. 2.6. The predominant location of GC clusters in the middle of AT spacers suggests that they may fulfil a stabilizing role.

1.5. The GC clusters

Figure 2.6. A schematic representation of the organization of a mitochondrial genome unit of yeast. Grey stretches represent genes, white stretches AT spacers, black bars GC clusters. (From Bernardi, 1979a).

1.6. The excision sites

The next step was the precise definition of the sequences involved in the excision process. The basic idea of the deletion model mentioned above was that the instability of the mitochondrial genome of yeast was due to the existence in each genome unit of a number of nucleotide sequences having enough homology to allow illegitimate, unequal recombination to take place. In this respect, clearly the GC clusters were at least as good candidates as the AT spacers.
Investigations on a number of spontaneous petite mutants showed that, indeed, most frequently the ends of the repeat units were formed by (CCGG, GGCC) clusters; less frequently, they appeared to correspond to GC-rich clusters and to AT spacer sequences (Faugeron-Fonty et al., 1979; Gaillard et al., 1980). For example, in the petite genomes of Fig. 2.7, excision of the repeat units occurred at, or very near, (CCGG, GGCC) clusters in the top four cases; and at GC clusters or AT spacers in the other two. Furthermore, it was shown that repeat units were organized in a perfect tandem (head-to-tail) fashion. Interestingly, secondary excision of simpler repeat units from the petite genomes originally derived from the parental wild-type genome appeared to take place at the same kind of sites used in the primary process, the end product being rather simple and stable petite genomes, such as those presented in Fig. 2.7. The spontaneous, cytoplasmic petite mutation should, therefore, be visualized as a cascade of excisions, which is slowed down, or stopped, only because sequences appropriate for excision are used up in the process, or because the DNA looping needed for recombination becomes impossible due to the short distance between potential recombination sites.

Such a simple excision mechanism was not found in petites induced by ethidium bromide, practically the only ones studied in other laboratories. These contained inverted restriction enzyme maps of the repeating units of the mitochondrial genomes of several spontaneous petite mutants. The molecular weights of the repeat units are indicated, along with the positions of *Hae* III (Δ), *Hpa* II (▼) and other restriction sites. In the case of a3B, five isolated *Hpa* II sites and *Hinf* II site are not shown. The broken lines indicate corresponding restriction sites in different repeat units. (From Bernardi, 1979a).
repetitions, multiple deletions, sequence rearrangements and, above all, were caused by much less specific excisions (Lewin et al., 1978). This is not surprising if one considers that the tremendous increase in petite formation upon mutagenization is accompanied by extensive genome fragmentation (Goldring et al., 1970) and that petites lacking mitochondrial DNA altogether (called \( \rho^0 \) petites in contrast to the \( \rho^- \) petites, which contain mitochondrial DNA) are frequently formed (Goldring et al., 1970; Nagley and Linnane, 1970). (By analogy, we called \( \phi^- \) the bleached mutants of \textit{Euglena gracilis} because, even if traditionally thought to be completely deprived of chloroplast DNA, they do contain a defective chloroplast genome, which is present at a very low copy number, and preferentially retains ribosomal RNA genes; Heizmann et al., 1981).

It was obvious that what was needed was detailed knowledge, at the nucleotide level, of the sequences involved in the excision of petite genomes. The simplest interpretation of the sequence data was that excision was due to a crossing-over process, and that the primary event in the spontaneous petite mutation was very similar to the excision of the lambda prophage from the \textit{E.coli} chromosome, or to the dissociation of a bacterial transposon from its host plasmid; in this case, the GC clusters and sequences in the AT spacers play the same role as the insertion sequences delimiting a bacterial transposon.

1.7. Genomes without genes

The first determination of the nucleotide sequence of the repeat unit of the mitochondrial genome of the spontaneous petite mutant \( a_{1/RZ1} \) (Gaillard and Bernardi, 1979; Fig. 2.8)

![Figure 2.8. Nucleotide sequence of the repeat unit of the mitochondrial genome of spontaneous petite mutant \( a_{1/RZ1} \) (see Fig. 2.7). A, B and C indicate GC-rich clusters; \( a \) and \( b \), a repeated decanucleotide and a repeated hexanucleotide, respectively; \( a', b', c' \) and \( d' \), palindromic sequences in the AT stretches. The restriction sites of \( \text{Hpa II} \), \( Mbo I \) and \( Mbo II \) are indicated by arrows; the recognition site of the latter enzyme is also indicated (From Bernardi, 1979a).](image-url)
and preceding work (Cosson and Tzagoloff, 1979; Macino and Tzagoloff, 1979) provided complementary confirmations of several previous results, namely that AT spacers are made up of short alternating and non-alternating AT sequences (Bernardi and Timasheff, 1970; Ehrlich et al., 1972) and contain direct and inverted repeated sequences and palindromes (Prunell and Bernardi, 1977); that GC-rich clusters are largely contiguous to CCGG sequences and to (CCGG, GGCC) clusters (Prunell and Bernardi, 1977); and that the latter are to some extent endowed with homology (Prunell and Bernardi, 1977). The genome of a_{1/1R/Z1} was of interest in three other respects: (i) it did not contain any gene, and was therefore a clear example of the lethality of the petite mutation as far as mitochondrial functions are concerned; incidentally, the surprising finding of genomes without genes can be taken as the best evidence that genetics and genomics are not the same thing; (ii) it replicates; in fact this is the only function left; since this genome was shown to be made up of a perfect tandem repetition of the basic unit (Faugeron-Fonty et al., 1979), the latter must contain a signal permitting the initiation of replication; (iii) it is excised, in all likelihood, from another petite genome, a_{1/1R/1} (Fig. 2.7), and not directly from the wild-type genome; the complete sequence of the parental petite genome a_{1/1R/1} (see Fig. 2.12) should, therefore, provide precise information as to the excision sites involved in the formation of a_{1/1R/Z1}.

To summarize, the first event in the spontaneous cytoplasmic petite mutation is the excision of a segment from one of the 25–50 mitochondrial genome units (Fig. 2.9) of a

Figure 2.9. Physical and genetical map of the mitochondrial genome unit of wild-type yeast (stlam A). Some restriction sites are indicated. Circled numbers indicate the location of ori sequences 1–8 (arrowheads point in the direction cluster C to cluster A; see Fig. 2.8). Black and dotted areas correspond to exons and introns of mitochondrial genes, respectively. Thin radial lines ending in small circles indicate tRNA genes. White areas correspond to long AT spacers embedding short GC clusters (Modified from de Zamaroczy et al., 1979).
wild-type cell; the excision is then followed by a tandem amplification process in which the excised DNA segment becomes the repeat unit of a defective genome unit which replicates within the parental wild-type cell and segregates into the buds of the latter; further segregation of these defective genome units in the progeny leads to the formation of petite mutants, whose mitochondrial genome is formed by identical defective units.
2.1. Excision and recombination

The sequences used in the excision process were investigated in a number of spontaneous petite genomes (Marotta et al., 1982; de Zamaroczy et al., 1983). In all cases, excision sequences were perfect direct repeats located in the AT spacers or in the GC clusters, as predicted (see Bernardi, 1979a). Indeed, sequences of 11–12 base pairs could be used as excision sequences; longer excision sequences were also used and even included canonical and surrogate origins of DNA replication (see the following section); when shorter sequences were used, they were flanked by regions of patchy homology. These results also indicated: (i) that the highest excision rates were associated with sequences that could form the most stable (longest and/or GC-richest) heteroduplexes; (ii) that the production of petite mutants depended not only upon the excision rate, but also upon the replication rate of the defective genome relative to the parental one, and upon its stability, namely its lack of susceptibility to further excisions (Marotta et al., 1982); both these parameters favored the production of petite mutants carrying mitochondrial genomes formed by short repeat units containing canonical ori sequences (see the following section); (iii) that the excision mechanism involved unequal crossing-over events within a genome unit and that this process probably was just a special case of the very active recombination processes taking place in the mitochondrial genomes of wild-type yeast cells (Fonty et al., 1978).

Indeed, when *S. cerevisiae* wild-type strains were crossed, recombinant mitochondrial genomes were present in 75% of the diploid clones. Such recombinant genomes had unit sizes different from, yet within 5% of, the parental ones and showed *Eco RI* and *Hind II + III* fragment patterns of parental types, two strong indications that both the gene complement and the gene order were very largely preserved in the progeny. Indeed, fragment patterns produced by *Hpa II* and *Hae III* were characterized by (i) fragments originating from the DNAs of both parents; and (ii) new fragments, *i.e.* fragments absent in either parent (Figs. 2.10 and 2.11). The new fragments appeared to arise from unequal crossing-over events occurring in the corresponding spacers of parental genetic units and usually had preferential locations in the genome.

These results provided the first evidence for physical recombination of mitochondrial DNA in crosses of wild-type yeast cells, indicated that recombination was very frequent in crosses and shed some light on mitochondrial segregation. They also provided examples of recombination phenomena in an interspersed system of unique and repetitive nucleotide sequences.
Figure 2.10. Scheme of the Hpa II band patterns on 0.5% agarose/2% polyacrylamide, 0.5% agarose/3% polyacrylamide, 6% polyacrylamide, as shown by the DNAs from parent strains A and B, and the diploid clones issued from the mass-mating experiment (1 1, 2.1, 4.1, 5.1, 8.1, 9.1). The molecular weights of the fragments are indicated (values in daltons are multiplied by $10^4$ in the left-hand pattern, by $10^5$ in the other two). Bands containing 1 or 2 (or more) fragments are indicated by a different thickness. This scheme only depicts the bands present in the best-resolved region of the gel and does not show the overlap region between different gels. Black bands are bands common to both parental strains; green and blue bands are bands only present in one of the two parents; red bands are bands differing in mobility or multiplicity from parental bands. Faint bands are indicated by broken lines (the faintest bands being indicated by 4 dashes, the others by 3). (From Fonty et al., 1978).

2.2. The canonical and the surrogate origins of replication of petite genomes

The mitochondrial genomes of the vast majority of spontaneous petites are exclusively derived from the tandem amplification of a DNA segment excised from any region of the parental wild-type genome (Faugeron-Fonty et al., 1979). Therefore, either the wild-type genome contains several origins of replication and at least one of them is present on the excised segment (Prunell and Bernardi, 1977), or sequences other than the origins of
replication of the wild-type genome are used as surrogate origins of replication. In fact, both situations have been found to occur, although with very different frequencies.

Considering that the first explanation was the more likely one, when we first sequenced (Gaillard and Bernardi, 1979) the repeat units of two petite genomes excised from the same region of the wild-type genome, we looked for a putative origin of replication in the segment shared by them and found a region characterized by two short GC clusters, A and B, flanking a palindromic AT sequence, p, and a short AT segment, s; and one long GC cluster, C, separated from B by a long AT segment, l (see Figs. 2.8 and 2.12). We also noticed (de Zamaroczy et al., 1981) that the potential secondary structure of the A-B region, the primary structure of cluster C and the general arrangement of the whole region were remarkably similar to those found in animal mitochondrial origins of replication (Crews et al., 1979; Gillum and Clayton, 1979; Fig. 2.13).

An ori sequence like the one just described was found in almost all the mitochondrial genomes of spontaneous petite mutants. Restriction mapping and hybridization of petite genomes on restriction fragments of wild-type genomes (de Zamaroczy et al., 1981; Bernardi, 1982b) provided evidence for the existence of eight ori sequences in the mitochondrial genome of wild-type cells. The primary structure of the ori sequences showed that they were extremely similar in sequence, particularly in their GC clusters (Fig. 2.12). All these canonical ori sequences have been precisely localized and oriented on the physical map of the genome (Fig. 2.9).

It should be noted: (i) that some ori sequences display one orientation and some the opposite one on the wild-type genome; (ii) that ori 2 and 7 as well as ori 3 and 4 are close to each other and tandemly oriented; (iii) that ori 4 is absent in a wild-type strain; (iv) that ori sequences containing the γ cluster were found only once (ori 4), or not at all (ori 6, ori 7) in extensive screenings of spontaneous petite genomes (see Table 2.1).

Ori° petites, lacking a canonical ori sequence, were also found, although very rarely. An investigation of the mitochondrial genomes of eight such ori° petites (Goursot et al., 1982) has revealed that their repeat units contain, instead of canonical ori sequences, one or more ori° or ori “surrogate” sequences. These ori° sequences are a subset of GC clusters char-
Figure 2.12. Primary structure of the ori sequences and their flanking regions. Regions of ori sequences are indicated by thick lines for GC clusters A, B and C and thin lines for AT stretches. The positions of GC clusters α in ori 4, β in ori 4 and 6, γ in ori 4, 6 and 7, and δ in ori 7 are given; r sequences are indicated by heavy line boxes; r and r sequences by broken line boxes. Other boxes indicate sequences la, lp, lg, ra, the excision sequences of the repeat units of petites a-1/1R/1/26, a-1/1R/14, a-3/1/33, a-3/1/5, and a-1/1R/2/1 and the initiation triplet of the open reading frames of ori 1, 2, 3 and 5. (From de Zamaroczy et al., 1983).

characterized by a potential secondary fold with two sequences ATAG and GGAG inserted in AT spacers; these sequences are followed by two AT base pairs, a GC stem (broken in the middle, and in most cases also near the base, by non-paired nucleotides) and a terminal loop (Fig. 2.14). This structure is reminiscent of that of GC clusters A and B from canonical ori sequences (Fig. 2.13). Like the latter, ori sequences are present in both orientations, are located in intergenic regions and can be used as excision sequences. Ori sequences are homologous with many other subsets of GC clusters (one of these subsets, the ori-like sequences, is shown in Fig. 2.14) some or all of which might perhaps also act as surrogate origins of replication, possibly still less efficiently than ori sequences.
35

Figure 2.13. A. Comparison of ori sequences of mitochondrial genomes from yeast (de Zamaroczy et al., 1981) and HeLa cells (Crews et al., 1979). Homology of potential secondary structures is found for the inverted repeats in the cluster A - cluster B region; arrows indicate the base changes found in this region in different petite genomes. Homology of primary structure is found for cluster C. B. Comparison of the two ori sequences; the arrows indicate the inverted repeats of the A-B region, the broken line corresponding to the looped-out sequence; bp, base pairs. (From de Zamaroczy et al., 1981).

Figure 2.14. Potential secondary structure of: A the ori sequences; and B the ori-like sequences. All sequences are drawn in the same orientation ATAG → GGAG. Double-headed arrows indicate base-exchanges, arrows pointing towards, or away from, the structure indicate insertions and deletions, respectively. Numbers indicate the ori, or ori-like, sequences presenting these changes. (From Goursot et al., 1982)

2.3. The replication of petite genomes and the phenomenon of suppressivity

A functional evidence that ori sequences are indeed involved in the replication of the mitochondrial genome came from crosses of spontaneous petites, characterized in their mitochondrial genome and their suppressivity, with wild-type cells (de Zamaroczy et al., 1979; 1981; Goursot et al., 1980). In the suppressivity test, petite mutants are crossed with
wild-type cells. The degree of suppressivity (or suppressiveness), namely the percentage of diploid petites in the progeny (which carry in every case the parental petite genome; Goursot et al., 1980; de Zamaroczy et al., 1981; Rayko et al., 1988) is determined by the replicative efficiency of the petite genome relative to the mitochondrial genome of the wild-type cells used in the cross (Bernardi et al., 1980a,b; Blanc and Dujon, 1980; de Zamaroczy et al., 1981). Crosses of spontaneous, highly suppressive petites having mitochondrial genomes formed by very short repeat units (400–900 base pairs) with wild-type cells produced diploids which harbored only the unaltered mitochondrial genomes of the parental petite, which was called supersuppressive petite (de Zamaroczy et al., 1979; Goursot et al., 1980). When petites with different degrees of suppressivity were used in the crosses, the genomes of diploid petite progeny had restriction maps identical to those of the parental haploid petites. Very few exceptions were found and these corresponded to new excision processes affecting one of the parental genomes.

There are two clear correlations between the ori sequence of the petite used in the cross and its degree of suppressivity. First, all other properties being comparable (namely, identical genetic background, the intact state of the ori sequence and the total amount of mitochondrial DNA per cell), the lower the overall density of ori sequences on the genome units, the lower the suppressivity. This can be called the repeat-unit-size rule (see Section 2.5). Second (see Table 2.1), partial or total deletion of the ori sequences, or their rearrangement, affect suppressivity: (i) Ori− petites, in which the ori sequence is partially deleted, show a decreased suppressivity relative to ori+ petites carrying intact ori sequences; the loss of cluster C with its flanking sequence has a much more dramatic effect than the loss of cluster A; (ii) orf petites, which show an inverted orientation of two ori sequences within the same repeat units (the latter having, in turn, an alternate inverted and tandem orientation) have a very low degree of suppressivity; (iii) ori° petites, which lack the ori sequence altogether but contain orf sequences instead, have low to minimal

| TABLE 2.1 |
| Replication and transcription of petite genomes

<table>
<thead>
<tr>
<th>Petites</th>
<th>Suppression</th>
<th>Transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Ori+ petites</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ori 1</td>
<td>&gt;95%</td>
<td>+</td>
</tr>
<tr>
<td>ori 2</td>
<td>&gt;95%</td>
<td>+</td>
</tr>
<tr>
<td>ori 3</td>
<td>85%</td>
<td>+</td>
</tr>
<tr>
<td>ori 4°</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ori 5</td>
<td>90%</td>
<td>+</td>
</tr>
<tr>
<td>ori 6°</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>ori 7°</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>B Ori− petites</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ori 1 A°</td>
<td>80%</td>
<td>+</td>
</tr>
<tr>
<td>ori 1 C°</td>
<td>n.d.</td>
<td>–</td>
</tr>
<tr>
<td>ori 3 C°</td>
<td>&lt;5%</td>
<td>–</td>
</tr>
<tr>
<td>C Orf petites</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a-15/4/110/3</td>
<td>~1%</td>
<td>–</td>
</tr>
<tr>
<td>a-3/1/B4°</td>
<td></td>
<td>–</td>
</tr>
<tr>
<td>D Ori° petites</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;5%</td>
<td>n.d.</td>
<td>–</td>
</tr>
</tbody>
</table>

a Modified from Baldacci and Bernardi (1982).
b Values found for petite genomes having repeat units ~900 (ori 1, 2) or 1800 (ori 3, 5) base pairs long.
c Ori 4 was only found once, ori 6 and 7 were never found alone in the extensive screenings of spontaneous petite genome.
d Diploid.
degrees of suppressivity. These results provide a molecular basis for a replicative advantage being the explanation for suppressivity.

The first direct demonstration that the active ori sequences are indeed origins of DNA replication, as previously postulated on the basis of compelling but indirect evidence, was provided by Baldacci et al. (1984). These authors showed that in both ori 1 and ori 5, nascent DNA chains using as template the strand containing sequence r (the "r strand") start at the r end of cluster C, are elongated towards sequence l, and follow an RNA primer starting at sequence r. Nascent DNA chains copied on the "non-r strand" start within cluster C, are elongated towards sequence r, and follow an RNA primer starting in sequence l just before cluster C. Ori 1 and 5 are, therefore, used as sites for RNA-primed bidirectional replication of mitochondrial DNA (see Fig. 2.15).

Figure 2.15. Initiation of DNA replication. Cluster C and sequences r and r') and the end of sequence l are shown. Arrows correspond to the position of starting points for RNA and DNA synthesis for each strand of the two ori sequences. (From Baldacci et al., 1984).

2.4. The ori sequences as transcription initiation sites

That ori sequences also act as transcription initiation sites is indicated by three results on petite genomes (Baldacci and Bernardi, 1982).

(i) Transcription initiation efficiency parallels replication efficiency. Petite genomes containing some canonical ori sequences (ori 1, 2, 5, and, to a lesser extent, ori 3) are transcribed very actively; others, containing ori 4, or deleted in their C clusters (but not those deleted in their A clusters), or lacking canonical ori sequences (ori° petites), are not (Table 2.1). Likewise, ori sequences containing a γ cluster (ori 4, 6 and 7) are probably not very efficient in DNA replication, as suggested by the fact that they are very rarely or never found in extensive screenings of spontaneous petites and may even be absent (ori 4) in some wild-type genomes (Faugeron-Fonty et al., 1984).

(ii) Transcription initiates next to the oligopyrimidine stretch of cluster C, at a sequence (Fig. 2.12) which is very similar to the transcription initiation sequences of rRNA genes (Oisinga and Tabak, 1982), and proceeds from cluster C to cluster A. As already mentioned, the insertion of cluster γ in the middle of this sequence (as in ori 4) or the loss of
cluster C (as in ori C− petite genomes) is accompanied by a loss of transcriptional activity. This suggests that ori 2 and ori 5 might be among the initiation sites used for transcribing the sense strand. ori 1 and ori 3 among those used for the transcription of the other strand. A small number of other sequences largely homologous to the transcription initiation sites of tRNA genes have also been found. These might play a role in the multipromotor transcription (Levens et al., 1981) of the mitochondrial genome of wild-type cells which had been postulated by Prunell and Bernardi (1977).

(iii) Since transcriptionally active ori sequences (see above) are present in both orientations on the wild-type genome, it is very likely that both strands are transcribed, although the non-sense strand appears to be transcribed less accurately, or more slowly. Hybridization experiments with separated DNA strands have identified the template strand used in transcription as the strand which contains the oligonucleotide stretch of cluster C. This conclusion supports previous independent evidence (Coruzzi et al., 1981; Baldacci and Bernardi, 1982; Beilharz et al., 1982) and puts the transcription of the mitochondrial genome of yeast in line with that of the animal mitochondrial genome. Similarly, replication might proceed unidirectionally from some ori sequences (possibly ori 2 and 5 for one strand, and ori 1 and 3 for the other).

2.5. The effect of flanking sequences on the efficiency of replication of petite genomes

The mitochondrial genomes of progenies from 26 crosses between 17 cytoplasmic, spontaneous, suppressive, ori+ petite mutants of S. cerevisiae were studied by electrophoresis of restriction fragments. Only parental genomes (or, occasionally, genomes derived from them by secondary excisions) were found in the progenies of the almost 500 diploids investigated; no evidence for intermolecular unequal mitochondrial recombination was detected. One of the parental genomes was always found to predominate over the other one, although to different extents in different crosses. This predominance appears to be due to a higher replication efficiency, which is correlated with a greater density of ori sequences on the mitochondrial genome (i.e. with a shorter repeat unit size of the latter).

Exceptions to this ‘repeat-unit-size rule’ (see Section 2.3) were found, however, even when the parental mitochondrial genomes carried the same ori sequence (compare petites b11 and b7, a10/1 and a10/20, for instance). This indicates that noncoding, intergenic sequences flanking ori sequences also play a role in the modulation of replication efficiency (see Fig. 2.16). Since in different petites such sequences differ in primary structure, size, and position relative to ori sequences, this modulation is likely to take place through an indirect effect on DNA and nucleoid structure (Rayko et al., 1988; see Fig. 2.21).
2.6. *The ori* petites 14 and 26

The effect of temperature on the replicative ability of the mitochondrial genomes was investigated on three isonuclear, spontaneous, cytoplasmic petite mutants of *S. cerevisiae*, al/1R/14, al/1R/1/26 and al/1R/Z1, which here will be referred to as petites 14, 26 and Z1, respectively (Goursot et al., 1988). The primary structure of the mitochondrial genomes of these petites had been previously determined (Gaillard et al., 1980; de Zamaroczy et al., 1981; 1983; 1984). Each one of these genomes arose from the excision of a DNA segment from the mitochondrial genome of yet another petite, al/1R/1 (see Fig. 2.17). As is the rule for spontaneous, cytoplasmic petites (Gaillard et al., 1980; de Zamaroczy et al., 1983), these excisions (i) took place by unequal recombination events involving pairs of short direct repeats (shown in Fig. 2.17) and (ii) were followed by tandem amplification of the excised segments, which became the repeat units of the newly formed genomes of the secondary petites Z1, 14 and 26. While the repeat unit of the parental petite genome is 884 bp long, the repeat units of petites Z1, 14 and 26 are 416, 392 and 398 bp long, respectively. In the latter three cases, each repeat unit corresponds to about 0.5% of the wild-type mitochondrial genome and only contains about 100 bp in addition to one *ori* sequence, *ori*1. The amount of mitochondrial DNA in these petites being roughly the same as in wild-type cells, these petites contain about 50 times (200/4) more *ori* sequences than the parental wild-type cells.

While each repeat unit of the mitochondrial genomes of petites al/1R/1 and Z1 contains a complete *ori* sequence, those of petites 26 and 14 contain *ori* sequences which had lost 11 and 27 bp, respectively, in the excision process. In the case of petite 26, the deletion comprises GC cluster A; in the case of petite 14, it comprises also part of the neighbouring sequence *p* (de Zamaroczy et al., 1981; 1984; Figs. 2.12 and 2.17). In the complete *ori*
Figure 2.17 The repeat unit of the mitochondrial genome of primary petite al/1R/l (see Gaillard et al., 1980) is 884 bp in size and contains an ori sequence, oriI. This comprises three GC clusters, A, B and C (heavy overlines). Cluster B is separated from cluster A by two short AT sequences p and s, and from cluster C by a long AT sequence f (broken overline). Cluster C is followed by sequence r (light overline) which comprises, on the strand shown, a nonanucleotide (heavy underline) where transcription of the RNA primer for the other nascent DNA chain starts before cluster C (arrow pointing to the right) (Baldacci et al., 1984). Sequence r is followed by a sequence ra which might pair with sequence la (within the l sequence) to form a hypothetical tertiary structure (de Zamaroczy et al., 1984: see Fig. 2.21). In the case of oril, sequence r is preceded by a duplication, sequence r. Excision sequences (e) of the secondary mitochondrial genomes from petites 14, 26 and Z1 (see Gaillard et al., 1980; de Zamaroczy et al., 1984) are indicated by boxes. (From Goursot et al., 1988).

sequences of ori+ petites al/1R/l and Z1, a secondary structure, the ‘A-B fold’, is possible (de Zamaroczy et al., 1981; 1984). This is a stem-and-loop structure, comprising 11 AT bp and 6 GC bp, which is formed by GC clusters A and B and the sequence in between (Figs. 2.17–2.19). In contrast, the formation of this structure is impossible in the ori- genomes of petites 14 and 26, which can form, however, ‘replacement folds’ made of pure AT (Fig. 2.19), endowed, therefore, with a lower thermodynamical stability (especially in the case of petite 14) compared to the A-B fold (de Zamaroczy et al., 1984).

The ori- mitochondrial genome of petites 14 and 26, which lack a normal feature of the ori sequence, the A-B fold (a feature which is perfectly conserved in all ori sequences; de

Figure 2.18. The repeat units of the mitochondrial genomes of petites Z1, 26 and 14. The oriI sequence, with its main elements, GC clusters A, B, C (black boxes) and sequence r (hatched box), is also represented. (From Goursot et al., 1988).
Figure 2.19. Potential secondary structure of the A-B fold of the ori sequence present in the repeat units of mitochondrial genomes from petites a1/1R/1 and Z1 (ori1) and of the replacement folds that can be formed in the ori1 sequence present in the mitochondrial genomes of petites 14 and 26. In the case of petite 14, the residual nucleotides from the partially deleted p stretch (pΔ) can generate a hairpin structure with nucleotides from the preceding repeat unit, but the stem, only formed by 13 A:T nucleotides, carries different terminal and side loops compared to the A-B fold. In the case of petite 26, the upper part of the stem and terminal loop are identical to those of the A-B fold, but the lower part is replaced by three A:T pairs (three nucleotides are derived from the preceding repeat unit). The sequence involved in the structure shown (GC clusters A and B, sequences p and s) are those indicated in Fig. 2.13. (From Goursot et al., 1988).

Zamaroczy et al., 1981; 1984), do exhibit lower replicative abilities compared to the ori+ mitochondrial genomes of petites a1/1R/1 and Z1 (as well as of all ori− petite genomes having repeat units shorter than about 1000 bp; Rayko et al., 1988). This can be judged by the suppressivity test (Ephrussi et al., 1955). Indeed, when crossed with wild-type cells, ori− petites 26 and 14 yield progenies comprising only 85% and 79% diploid petites, respectively, instead of 99% produced by crosses involving ori+ petites a1/1R/1 and Z1 (Table 2.2). Incidentally, much more dramatic decreases in suppressivity are associated with petite genomes deleted in the crucial C cluster region (de Zamaroczy et al., 1981), or lacking ori sequences altogether; as already mentioned, these latter ori− petite genomes could replicate by making use of surrogate origins of replication, namely of ori− sequences.

**TABLE 2.2**
Suppressivity of petites Z1, 26 and 14 at different temperaturesa

<table>
<thead>
<tr>
<th>Temperature b (°C)</th>
<th>Petites:</th>
<th>Suppressivity (%)c</th>
</tr>
</thead>
<tbody>
<tr>
<td>23°</td>
<td>Z1</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>90</td>
</tr>
<tr>
<td>28°</td>
<td></td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>79</td>
</tr>
<tr>
<td>33°</td>
<td></td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>45</td>
</tr>
</tbody>
</table>

a From Goursot et al. (1988)
b Temperatures indicated were applied to precultures and to cultures used in crossings and in the incubation of diploids. Precultures in stationary phase were diluted ten times upon setting up the cultures used in crossings.
c Suppressivity was tested by crossing petite strains (MATa, ade1) against the wild-type strain B (MATα, trpl, his1). The percentage of petite/total diploid colonies was determined by counting about 1000 colonies on minimal agar.
2.7. Temperature and the replicative ability of ori<sup>−</sup> petites 14 and 26

In view of the presumed lower thermodynamical stability of replacement folds compared to the A-B fold, crosses of petites 14 and 26 with wild-type cells were also performed at both a higher (33°C) and a lower (23°C) temperature than the standard one, 28°C. The results of Table 2.2 indicate that such temperatures decreased and increased, respectively, the replicative ability of the two mutants, whereas that of the control ori<sup>+</sup> petite Z1 did not show any significant change between these temperatures. The decrease of the replicative ability was stronger for petite 14 than for petite 26, as expected from the presumably lower thermodynamical stability of the replacement fold of the former. In the first case, the proportion of wild-type colonies obtained at 23°C and 33°C, respectively, was 10% and 55%, in the second 11% and 42%. The changes in replicative ability just described are immediate and reversible.

The results presented indicate that an environmental factor, temperature, can reversibly affect the replicative ability of a genome by altering its secondary (and possibly, its tertiary) structure. Indeed, (i) these changes cannot be ascribed to enzymes involved in DNA replication, as in temperature-sensitive mutants, since the petites discussed here are iso-nuclear, and lack mitochondrial protein synthesis, like all petites; (ii) the different effects of temperature on the replicative ability of petites Z1, 14 and 26 show an excellent correlation with those expected from the secondary structures of the postulated A-B fold and replacement folds (de Zamaroczy et al., 1981; 1984), an effect on tertiary structures being also possible.

It should be pointed out that the repeat unit of petite Z1 extends 80 bp to the left of cluster A and only 40 bp to the right of sequence r, whereas those of petites 26 and 14 extend 115 bp to the right of sequence r (Fig. 2.18); in all cases, however, these extensions to the left of ori sequences are just made of AT spacer. Effects of flanking regions of ori sequences on the replicative ability of the latter are known (Rayko et al., 1988; see Section 2.5), but they are small compared with the different suppressivities exhibited by petites 14 and 26 relative to petite Z1. There is, therefore, no doubt that these differences are due to the deletions in the ori sequences of petites 14 and 26.

Interestingly, the conclusion that DNA secondary structure is required for ori activity in vivo was also reached for bacteriophage G4, where a strong temperature-dependent impairment of replication was found after introducing by site-directed mutagenesis point mutations which destabilize intra-strand base-pairing in the ori sequence (Lambert et al., 1987).

From a general viewpoint, our results indicate the existence of a novel type of environment-genome interaction, in which reversible changes in higher-order DNA structures are induced with profound consequences on a basic genome function, such as replication. These changes consist in genome transconformations which, although non-inheritable as such, can be maintained for many generations in the presence of the appropriate environmental condition. Genome transconformations can provide, therefore, strong selective advantages or disadvantages and play an important role in evolution, in addition to classical mutations, which involve changes in the primary structure of DNA. Moreover, similar phenomena might (i) be induced by other environmental factors; (ii) affect other genome functions (e.g., transcription); and also (iii) be operative in other organisms.
CHAPTER 3

The organization and evolution of the mitochondrial genome of yeast

3.1. The organization of the mitochondrial genome of yeast

The organization of the mitochondrial genome of yeast, as it has emerged from our work, is similar to that of the nuclear genome of eukaryotes in that (i) coding sequences are interspersed with non-coding sequences, and (ii) the non-coding sequences are similar to the interspersed repeated sequences and the fold-back sequences, identical or similar sequences being present in many copies in the genome; moreover, (iii) the genome contains two genes, coxl (cytochrome oxidase sub-unit 1) and cyt b (apocytochrome b), that comprise several introns (in some strains), some of which are translated into maturases, reverse transcriptases or site-specific endonucleases; the 21S RNA gene also may comprise an intron (see Foury et al., 1998, for a review).

As already mentioned (see Chapter 1, Section 1.1), these genome features are not at all unique to the mitochondrial genome of yeast. Indeed, we have shown the presence of non-coding AT spacers in the mitochondrial genome of *Ustilago cynodontis* (Mery-Drugeon et al., 1981), a fungus belonging to the class of Basidiomycetes (in contrast of *S. cerevisiae* which is an Ascomycete), as well as in both the mitochondrial and chloroplast genomes of *Euglena gracilis*, a unicellular flagellate (Stutz and Bernardi, 1972; Fonty et al., 1975; Schmitt et al., 1981; Heizmann et al., 1981). A similar situation was also found in the mitochondrial genomes of other yeasts, *Torulopsis glabrata*, *Brettanomyces anomalus*, and *Kloeckera africana* (Clark-Walker and McArthur, 1978). Interestingly, another chloroplast DNA, from a higher plant, *Spinacia oleracea*, which is characterized by a higher GC level, 36.5%, also revealed a high degree of heterogeneity, 30% of the genome being only 22% GC and 10% of the genome being higher than 60% GC (Schmitt et al., 1981).

These results stress an important point, that mitochondrial genomes (to limit our discussion here to only one of the two organelle genomes) may range from the extremely compact structures that characterize not only most animal mitochondrial genomes, but also the mitochondrial genomes from some unicellular organisms (e.g., *Schizosaccharomyces pombe*), to structures in which non-coding sequences form the majority of the genome, the best studied case being that of the mitochondrial genome of yeast. The monophyletic origin of the mitochondrial genome from a prokaryotic ancestor then raises two problems concerning the mechanism of formation and the biological role of non-coding sequences in the large mitochondrial genomes of yeasts.

Before we move to these subjects in the following sections, it should be stressed that the organization of the mitochondrial genome of yeast points to the fact that in eukaryotic genetic systems, where so much of the DNA is non-coding, there is a real need for a molecular approach, because approaches based on classical genetics or on the study of gene products suffer from serious intrinsic limitations and are unable to provide an overall
picture of the genome. Incidentally, this organization disposed of a series of ideas, which were promoted for many years, and can be summarized as follows: (i) that the mitochondrial genome of yeast had a unique nucleotide sequence, namely a sequence lacking internal repetitions; this view had its origin in a misunderstanding of the renaturation kinetics of mitochondrial DNA; (ii) that the mitochondrial genome of yeast had, as a consequence, an informational content five times larger than that of animal mitochondrial genomes (which have a unit size of only 17 kb); (iii) that the unit size of the mitochondrial genome decreased in the evolution from unicellular organisms to animals. In conclusion, the mitochondrial genome is a useful, simple model for the nuclear genome.

3.2. The evolutionary origin of ori sequences

The extremely high homology of the eight canonical ori sequences indicates that they arose as the result of duplication and translocation events. More precisely, we proposed (see Fig. 2.20) that the canonical ori sequences derive from a primitive ori sequence (probably made of only a monomeric cluster C and its flanking sequences r* and r) through (i) a series of duplications and inversions generating clusters A and B; and (ii) an expansion process producing the AT stretches of ori sequences. It is possible that the ori sequences are folded in a tertiary structure, as in the hypothetical model of Fig. 2.21.
3.3. The evolutionary origin of the GC clusters

We studied the primary and secondary structures, the location and the orientation of the 196 GC clusters present in the 90% of the mitochondrial genome of *S. cerevisiae* which had been sequenced by 1986 (see de Zamaroczy and Bernardi, 1986b; the later completion of the sequence by Foury et al., 1998, did not change anything substantial in what follows). We found that (i) the vast majority of GC clusters is located in intergenic sequences (including *ori* sequences, intergenic open reading frames and the *var 1* gene; see next section) and in intronic closed reading frames (CRF’s); (ii) most of them can be folded into stem-and-loop structures; (iii) both orientations are equally frequent. The primary structures of GC clusters permit to group them into eight families, seven of which are clearly related to the family formed by clusters A, B and C of the *ori* sequences. Most GC clusters apparently originated from primary clusters also derived from the primitive *ori* sequence in the course of its evolution towards the present *ori* sequences (de Zamaroczy and Bernardi, 1986a).

3.4. The evolutionary origin of the AT spacers and the var 1 gene

Intergenic sequences represent 63% of the mitochondrial ‘long’ (85 kb) genome of *S. cerevisiae*. They comprise 170–200 AT spacers that correspond to 47% of the genome.
and are separated from each other by GC clusters, ORFs, ori sequences, as well as by protein-coding genes. Intergenic AT spacers have an average size of 190 bp, and a GC level of 5%; they are formed by short (20–30 nt on the average) A/T stretches separated by mono- to tri-C/G. An analysis of the primary structures of intergenic AT spacers has shown that they are characterized by an extremely high level of short sequence repetitiveness and by a characteristic sequence pattern; the frequencies of A/T isostichs (oligonucleotides having the same size) conspicuously deviate from statistical expectations, and exponentially decrease when their \((AT+TA)/(AA+TT)\) ratio, \(R\), decreases. A situation essentially identical was found in the AT spacers of the small mitochondrial genome (19 kb) of \(T. glabrata\). The sequence features of the AT spacers indicate that they were built in evolution by an expansion process mainly involving rounds of duplication, inversion and translocation events which affected an initial oligodeoxynucleotide (characterized by a particular \(R\) ratio) and the sequences derived from it. In turn, the initial oligodeoxynucleotide appears to have arisen from an ancestral promoter-replicator sequence which was at the origin of the nonanucleotide promoters present in the mitochondrial genomes of several yeasts (de Zamaroczy and Bernardi, 1987).

Common sequence patterns indicate that the AT spacers so formed gave rise to the \(var\) gene (by linking and phasing of short ORFs), coding for the only mitochondrially encoded protein of the large ribosome sub-unit, to the DNA stretches corresponding to the untranslated mRNA sequences and to the central stretches of ori sequences. The case of \(var\) gene (Hudspeth et al., 1982) is of special interest in connection with the idea of the expansion of ori sequences. This gene is 10% GC and contains a 46 bp GC cluster accounting for 38% of total GC. Its similarity with spacer-cluster sequences is so striking that it suggests that this gene arose from an intergenic sequence only recently. Incidentally, \(var\) provided the first example of the generation of coding from non-coding sequences (de Zamaroczy and Bernardi, 1987), a subject of great current interest (see Jordan et al., 2003, and Part 6, Section 2.3).

### 3.5. The non-coding sequences: evolutionary origin and biological role

To sum up the preceding three sections, as far as the intergenic sequences are concerned, it is conceivable that they were derived from ori sequences by the expansion process proposed by Bernardi (1982a) on the basis of sequence comparisons (Bernardi and Bernardi, 1980). This might have taken place through three different mechanisms, all of which are likely to have played a role. First, a slippage of the replicase could occur at the ori sequences; this is a well-known phenomenon first studied in the reiterative replication of poly (dAT:dAT) by DNA polymerase I of \(E. coli\) (Kornberg et al., 1964). A second mechanism is unequal crossing-over: evidence for the high frequency of such a phenomenon in mitochondrial recombination is available (Fonty et al., 1978). A third mechanism is insertion. Almost all GC clusters are inserted in AT spacers; some rare ones are inserted in AT-rich regions of rRNA genes (Sor and Fukuhara, 1982) and even in a protein-coding gene, \(var\) (see above). Interestingly, these insertions are not only transcribed but, in the case of \(var\), also translated.

Another interesting result (Bernardi, 1982a; de Zamaroczy and Bernardi, 1986a) was
that the closed reading frames (CRF) of the intervening sequences of \textit{oxt} 3 and \textit{cob} genes share all the features of intergenic non-coding sequences. Indeed, when the relative amounts of di- to hexa-nucleotides were compared to those from random sequences having the same sizes and compositions, they were found to exhibit the same deviations as the intergenic noncoding sequences of the mitochondrial genome. In contrast, intronic open reading frames (ORFs) showed oligonucleotide patterns which were generally quite distinct from those of CRFs, although some similarities could be detected in some cases (especially for \textit{aI5}o). The mitochondrial introns of yeast, therefore, are endowed with a mosaic structure, in which CRFs derive from mitochondrial intergenic sequences, whereas ORFs have a different origin (indicated as exogenous by other evidences).

It is evident that regulatory sequences acting as promotor, operators, sites for the initiation of replication, and sites involved in the processing of transcripts are present in the non-coding sequences of yeast mitochondrial DNA. Another function of the non-coding sequences is also well documented and has to do with illegitimate, unequal recombination. The excision of the spontaneous petite genomes just described is an example of these extragenic recombinational events. The same basic mechanism appears, however, to be more general and to account for: (i) the divergence of the mitochondrial genomes of wild-type yeast cells; it has been shown (Bernardi et al., 1975; Prunell et al., 1977) that different strains have mitochondrial genomes differing in the length of AT spacers, apparently the result of unequal crossing-overs in the sequences of spacers; (ii) similar changes in the mitochondrial genomes of the progeny arising from crosses of different wild-type strains (Fonty et al., 1978).

In conclusion, the evidence available at the present time appears to support the idea that the complex sequence organization of the mitochondrial genome of yeast corresponds to the needs of very active and finely regulated replication, transcription, and recombination processes. This seems to be achieved at the price of an exceptional genomic instability.

Twenty years ago, this genome instability was discussed as follows (Bernardi, 1983): The non-coding sequences of the mitochondrial genome of yeast are the source of three disadvantages for the genome. The main one is that the abundant direct repeats that they contain are potential excision sequences; as such, they are responsible for the extreme instability of the genome. The other two are that they increase replication time and energy expenditure. These disadvantages would quickly change wild-type yeast cells into suppressive and neutral petite mutants, if this intracellular selection were not counterbalanced by an intercellular selection in which the faster growing respiratory-competent wild-type cells compete out the respiratory-deficient petite mutants (incidentally, this accounts for the fact that in nature only wild-type yeast cells are found). Even if the disadvantages associated with the non-coding sequences do not lead, therefore, to the elimination of the mitochondrial genome, they should at least lead to the elimination of the non-coding sequences themselves. We know, however, that, although \textit{S. cerevisiae} strains exist which lack a number of intervening sequences (introns) and also \textit{ori} 4, in general non-coding sequences tend to be largely conserved. This indicates that the removal of non-coding sequences is selectively disadvantageous or, in other words, that non-coding sequences provide selective advantages which compensate for the disadvantages associated with them. This obviously raises the question of the nature of these advantages, namely of the physiological roles played by non-coding sequences.
It is clear from the genome map of the mitochondrial genome of yeast that the deletion of intergenic sequences, where most excision sequences used in the spontaneous mutation are located (de Zamaroczy et al., 1983), will frequently remove canonical ori sequences from the wild-type genome. Even if a wild-type genome lacking ori 4 has been found, it is evident that in general such elimination will affect replication and also transcription. There is therefore a selective advantage in keeping ori sequences in the wild-type genome. As far as non-coding sequences outside ori sequences are concerned, it should first of all be stressed that the expansion process does not propagate non-sense sequences, but propagates instead sequences which have been highly selected and conserved in evolution and whose primary role is to interact specifically with enzymes involved in DNA replication and transcription. Thus, the expansion of ori sequences leads to the propagation of potential regulatory signals, which may be used in the regulation of gene expression, in repression (anaerobiosis or glucose can shut off transcription) and derepression, in the processing of primary transcripts, and in the regulation of nucleo-mitochondrial interactions. Another physiological role of non-coding sequences concerns recombination, since evidence exists (Fonty et al., 1978) that repeated and palindromic non-coding sequences are involved in mitochondrial site-specific recombination. Finally, it should be recalled that some non-coding sequences appear to be inserted into transcribed genes or even be transformed into genes. In summary, a number of physiological roles apparently provide selective advantages compensating for the disadvantages inherent in the very existence of non-coding sequences. What we know about the non-coding sequences of the mitochondrial genome of yeast suggests that their conservation in the genome is due to selective advantages associated with their physiological roles; these sequences, or at least their majority, cannot, therefore, be considered “selfish DNA sequences” (sensu Doolittle and Sapienza, 1980; Orgel and Crick, 1980).

A final point on which the mitochondrial genome of yeast is relevant to the “selfish DNA” issue is the occurrence of functionless genomes in suppressive petites. Many of these genomes contain no gene, and yet replication, transcription and even transcript processing may still go on. In nature, as already pointed out, these genomes rapidly disappear since petites are competed out by the faster growing wild-type cells. Many of these genomes have such a replicative advantage over wild-type genomes that they could spread out through crosses with wild-type cells, if haploid. This does not occur in nature because parental wild-type cells and the derived petites have the same mating type. When isolated from competition with wild-type cells in the laboratory, however, petites not only survive, but frequently end up with very stable genomes which are the result of a selection on the basis of replication efficiency. These “genomes without genes” are practically made up of repeat units containing barely more than an ori sequence; replication is most efficient, the corresponding petites being supersuppressive, and transcription is only preserved because of the role played in replication. In other words, functionless genomes like the mitochondrial genomes of suppressive petites not only can exist and be quite stable, but they undergo a selection favoring those which are closest to the ultimate situation of being just a set of replication origins; this in vivo selection is very much the same found in the Qβ-replicase in vitro system by Mills et al. (1967). These “selfish genomes”, exemplifying primordial self-replicating systems, will however be lost in the long run, when mitochondrial or nuclear mutations will inactivate the initiation of replication.
Part 3

The organization of the vertebrate genome
Chapter 1

Isochores and isochore families

1.1. The fractionation of the bovine genome

The success met by the Cs$_2$SO$_4$/Ag$^+$ ultracentrifugation in separating satellite DNAs from mouse and guinea pig (Corneo et al., 1968; see Part 1, Section 1.3) encouraged us, at the beginning of the 1970's, to approach the fractionation of another mammalian DNA, bovine DNA, which was known (Corneo et al., 1970) to comprise two satellites (at that time calf thymus DNA was the standard mammalian DNA because it could be easily prepared in gram amounts). This allowed us (Filipski et al., 1973) to resolve the bovine genome into seven DNA components (see Fig. 3.1), which comprised: (i) four satellite DNAs, classified as such on the basis of their narrow CsCl bands, sharp melting curves and fast reannealing properties; these 1.705, 1.710, 1.714 and 1.723 satellite components (DNA components are indicated by their modal buoyant density in CsCl) represented 4%, 1.5%, 7% and 1.5% of the DNA, respectively; the last two were the only known satellites at the time of our work; note that, at this point, only four satellite DNAs were resolved from the major components, and that the separation of the remaining four satellite DNAs required

![Figure 3.1](image)

Figure 3.1. A A histogram of the DNA components of the bovine genome. The height of each bar is proportional to the percentage of each component in DNA; solid bars correspond to the sharp-melting, open bars to the broad-melting components (see B). The broken line is an enlarged band profile of calf DNA in CsCl density gradient. B. Absorbance-temperature profiles for calf thymus DNA components in 0.1 xSSC (standard saline citrate). Compare the very sharp transitions of 1.705, 1.714 and 1.723 g/cm$^3$ satellites with the broader ones of 1.697, 1.704 and 1.709 g/cm$^3$ major DNA components. (From Filipski et al., 1973)
a combination of different density gradient centrifugations (see below); and (ii) three major components, 1.697, 1.704 and 1.709, forming about 50%, 25% and 10% of the genome, respectively. Moreover, a number of minor components, formed together 4% of the genome; two of them, 1.719 and 1.699, were likely to correspond to ribosomal and mitochondrial DNAs, respectively (see Table 3.1).

Further work, in which we used both Ag⁺ and another sequence-specific DNA ligand, BAMD, bis(acetato-mercuri-methyl)dioxane (Cortadas et al., 1977; Macaya et al., 1978; Kopecka et al., 1978; Meunier-Rotival et al., 1979), led to a complete picture of bovine satellite and ribosomal DNAs. Indeed, these investigations (i) precisely defined the relative amounts and buoyant densities of both satellite and minor components and revealed that as much as 23% of the bovine genome was formed by eight GC-rich satellite DNAs; (ii) showed that, as a consequence, the standard deviation of the main CsCl band of bovine DNA had been overestimated by Sueoka (1959), a mistake justified by the fact that the work was performed at a time when the very existence of satellite DNAs was not yet known (see also Elton, 1974); and (iii) stressed the much higher resolving power of density gradient centrifugation over the reassociation approach.

As far as the last point is concerned, while the reassociation approach could only resolve four classes of sequences, non-repetitive, intermediate, fast and very fast (Britten and Smith, 1968; Davidson and Britten, 1973), density gradient centrifugation could resolve (Macaya et al., 1978) four major, eleven minor (originally defined as DNA components representing less than 3% of the genome) and eight satellite components (see Table 3.1). Moreover, a comparison of the two sets of data indicated that most of the bovine satellite sequences reassociate in the intermediate class. This demonstrated that not all satellites are fast-reassociating, as originally suggested on the basis of the results obtained with mouse satellite DNA. The slow reassociation of the satellite DNA from the bovine genome was

<table>
<thead>
<tr>
<th>Method of detection</th>
<th>Type of component</th>
<th>Number</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Renaturation kinetics</td>
<td>Non-repetitive</td>
<td>1</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>Intermediate</td>
<td>1</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Fast</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Very fast</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>4</td>
<td>98</td>
</tr>
<tr>
<td>Density gradient centrifugation</td>
<td>Major</td>
<td>4</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>Minor</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Satellite</td>
<td>8</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>23</td>
<td>100</td>
</tr>
</tbody>
</table>

*From Macaya et al. (1978). In this paper the lightest major component of Filipski et al. (1973) had already been resolved into two components. Renaturation kinetics data are from Britten and Smith (1968).*
due to the sequence divergence of the short repeats forming them, as shown by the sequencing of the major, 1.715 g/cm$^3$, satellite (Gaillard et al., 1981). In addition, this sequence and a restriction enzyme analysis allowed us to conclude that all eight satellite DNAs had a common evolutionary origin because they shared similar short basic repeats (Macaya et al., 1978; Kopecka et al., 1978). This conclusion was later confirmed by direct sequencing work (Pech et al., 1979). Further work demonstrated that, indeed, the 1.719 component was ribosomal DNA (Meunier-Rotival et al., 1979).

The demonstration of three well-defined major components by Filipski et al. (1973) forming the main band (Fig. 3.1) was an important novel finding, since it disproved the then generally accepted view that the bulk of the genome of higher organisms was formed by DNA molecules showing a continuous variation in GC level.

1.2. The fractionation of eukaryotic main-band DNAs

The fact that the three major components detected by Filipski et al. (1973) accounted for most or all of the main-band DNA of the bovine genome raised the possibility that they might be shared by all mammalian DNAs. Moreover, if the discontinuous compositional pattern was common to all mammalian genomes, one should explore more phyla to detect how far the compositional heterogeneity seen in the bovine genome could be found. Base composition, sedimentation coefficient, modal and mean buoyant density in CsCl were, therefore, investigated for DNAs from 25 eukaryotes ranging from human to yeast (see Table 3.2 and Fig. 3.2). This information (Thiery et al., 1976) was subsequently extended to many more vertebrates (see Part 4). Moreover, four mammalian and two amphibian DNAs were fractionated by preparative Cs$_2$SO$_4$/Ag$^+$ density gradient centrifugation and the CsCl band profile of each preparative fraction was analyzed in terms of Gaussian curves (Fig. 3.3), to allow assessing the relative amount and the buoyant density of each DNA component (Fig. 3.4). This more laborious, but more informative, approach was later extended to several fishes (see Parts 4 and 12) and to two reptiles (see Part 5).

The main conclusions of Thiery et al. (1976) were the following: (i) all eleven mammalian DNAs analyzed exhibited the three major components 1.697, 1.704 and 1.709 g/cm$^3$, first observed in the bovine genome; similar components appeared in avian genomes and were, likewise, responsible for the trailing of the CsCl bands on the heavy, GC-rich, side; (ii) the DNAs of reptiles, amphibians and fishes showed a much lower skewness to the heavy side of their CsCl bands; (iii) essentially symmetrical bands in CsCl, neglecting satellite bands, were displayed by DNAs from three invertebrates (two echinoderms and Drosophila) and from three unicellular eukaryotes, S. cerevisiae, E. gracilis, T. pyriformis (see Table 3.2).

The work of Thiery et al. (1976), besides collecting a number of new observations on minor and satellite components, led to the first recognition of major phylogenetic differences at the macromolecular level in the organization of eukaryotic genomes. Minor differences in satellite DNAs may also occur in rather closely related species because even a single nucleotide change in a short basic repeat (for instance, a heptanucleotide) may lead to considerable changes in the buoyant density and compositional properties of the satellite.
<table>
<thead>
<tr>
<th>DNA source</th>
<th>Tissue</th>
<th>S$_{20,w}$</th>
<th>G+C (%)</th>
<th>ρ$_0$ (%)</th>
<th>&lt;ρ&gt; (%)</th>
<th>ρ$_o$ (%) mg/cm$^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ANIMALIA</strong>&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CHORDATA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MAMMALIA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Homo sapiens</em> (human)</td>
<td>Leucocytes</td>
<td>25</td>
<td>1.6985</td>
<td>1.7008</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Placenta</td>
<td>22.3</td>
<td>1.6990</td>
<td>1.7010</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td><em>Bos taurus</em> (calf)</td>
<td>Liver</td>
<td>25.2</td>
<td>1.7000</td>
<td>1.7039</td>
<td>3.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thymus</td>
<td>26</td>
<td>1.6997</td>
<td>1.7033</td>
<td>3.6</td>
<td></td>
</tr>
<tr>
<td><em>Fels domestus</em> (cat)</td>
<td>Liver</td>
<td>32.7</td>
<td>1.7002</td>
<td>1.7031</td>
<td>2.9</td>
<td></td>
</tr>
<tr>
<td><em>Canis familiaris</em> (dog)</td>
<td>Liver</td>
<td>31.3</td>
<td>1.7005</td>
<td>1.7040</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td><em>Cavia porcellus</em> (guinea pig)</td>
<td>Liver</td>
<td>24.8</td>
<td>1.6982</td>
<td>1.7010</td>
<td>2.8</td>
<td></td>
</tr>
<tr>
<td><em>Cricetus norvegicus</em> (Chinese hamster)</td>
<td>Liver</td>
<td>22.6</td>
<td>1.7000</td>
<td>1.7015</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td><em>Rattus</em> sp (rat)</td>
<td>Liver</td>
<td>27.6</td>
<td>1.7006</td>
<td>1.7021</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td><em>Mus musculus</em> (mouse)</td>
<td>Liver</td>
<td>24.5</td>
<td>1.7007</td>
<td>1.7008</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thymus</td>
<td>21.6</td>
<td>1.7008</td>
<td>1.7016</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td><em>Glis glis</em> (garden dormouse)</td>
<td>Liver</td>
<td>35.8</td>
<td>1.6991</td>
<td>1.7004</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td><em>Oryctolagus cuniculus</em> (rabbit)</td>
<td>Liver</td>
<td>18.6</td>
<td>1.6999</td>
<td>1.7033</td>
<td>3.4</td>
<td></td>
</tr>
<tr>
<td><em>Erinaceus</em> sp. (hedgehog)</td>
<td>Thymus</td>
<td>32.8</td>
<td>1.7004</td>
<td>1.7035</td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td><strong>AVES</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Larus argentatus</em> (sea-gull)</td>
<td>Liver</td>
<td>23.7</td>
<td>1.6998</td>
<td>1.7031</td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td><em>Gallus</em> sp. (chicken)</td>
<td>Embryo</td>
<td>20</td>
<td>1.7001</td>
<td>1.7031</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td><strong>REPTILIA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Iguana iguana</em></td>
<td>Erythrocytes</td>
<td>20.5</td>
<td>43.9</td>
<td>1.7015</td>
<td>1.7022</td>
<td>0.7</td>
</tr>
<tr>
<td><em>Testudo graeca</em></td>
<td>Liver</td>
<td>25.1</td>
<td>46</td>
<td>1.7027</td>
<td>1.7042</td>
<td>1.5</td>
</tr>
<tr>
<td><strong>AMPHIBIA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Pleurodeles waltli</em></td>
<td>Testes</td>
<td>27.1</td>
<td>46.5</td>
<td>1.7041</td>
<td>1.7047</td>
<td>0.6</td>
</tr>
<tr>
<td><em>Xenopus laevis</em></td>
<td>Erythrocytes</td>
<td>35</td>
<td>40.9</td>
<td>1.6991</td>
<td>1.6979</td>
<td>0.6</td>
</tr>
<tr>
<td><strong>PISCES</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Opsanus tau</em> (toadfish)</td>
<td>Liver</td>
<td>20</td>
<td>42</td>
<td>1.7002</td>
<td>1.7004</td>
<td>0.2</td>
</tr>
<tr>
<td><em>Salmo trideus</em> (salmon)</td>
<td>Testes</td>
<td>20.5</td>
<td>43.5</td>
<td>1.7028</td>
<td>1.7035</td>
<td>0.7</td>
</tr>
<tr>
<td><strong>ECHINODERMATA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ECHINOIDEA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Strongylocentrotus purpuratus</em></td>
<td>Sperm</td>
<td>17</td>
<td>36.6</td>
<td>1.6989</td>
<td>1.6995</td>
<td>0.6</td>
</tr>
<tr>
<td><em>Paracentrotus lividus</em></td>
<td>Sperm</td>
<td>19</td>
<td>35.4</td>
<td>1.6972</td>
<td>1.6974</td>
<td>0.2</td>
</tr>
<tr>
<td><strong>ARTHROPODA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>INSECTA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>Embryo</td>
<td>27.4</td>
<td>39.1</td>
<td>1.7025</td>
<td>1.7014</td>
<td>−1.1</td>
</tr>
</tbody>
</table>
### TABLE 3.2
Properties of eukaryotic DNAs\(^a\) (continued)

<table>
<thead>
<tr>
<th>DNA source</th>
<th>Tissue</th>
<th>(S_{20w})</th>
<th>G+C (%)</th>
<th>(\rho_0) (b) (%)</th>
<th>(&lt;\rho&gt;) (b) (%)</th>
<th>(&lt;\rho&gt; - \rho_0) (b) (mg/cm(^3))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FUNGI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASCOMYCOTA</td>
<td>Total</td>
<td>28</td>
<td>37.1</td>
<td>1.6995</td>
<td>1.6984</td>
<td>-1.1</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td></td>
<td>40(c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PROTISTA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EUGLENOPHYTA</td>
<td>Nuclear</td>
<td>16</td>
<td>49.9</td>
<td>1.7080</td>
<td>1.7080</td>
<td>0</td>
</tr>
<tr>
<td>Euglena gracilis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CILIOPHORA</td>
<td>Macro nuclei</td>
<td>17.7</td>
<td>29.7</td>
<td>1.6910</td>
<td>1.6910</td>
<td>0</td>
</tr>
<tr>
<td>Tetrahymena pyriformis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MONERA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EUBACTERIAE</td>
<td>Total</td>
<td>34</td>
<td>51</td>
<td>1.7105</td>
<td>1.7101</td>
<td>-0.4</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(a\) From Thiery et al. (1976).

\(b\) The error on the \(\rho_0\) values is \(\pm 0.0005\) g/cm\(^3\); that on the \(<\rho>\) values is \(\pm 0.0002\) g/cm\(^3\).

\(c\) Systematic nomenclature according to Whittaker (1969).

\(d\) \(<\rho>\) and \(<\rho> - \rho_0\) values calculated for main band DNA, not taking into account the contribution of apparent satellites.

\(e\) DNA free of mitochondrial DNA and of the 1.705 g/cm\(^3\) satellite (ribosomal DNA).

---

**Figure 3.2.** CsCl profiles of DNAs A from mammals and B from birds, reptiles, amphibians, fishes, invertebrates and unicellular eukaryotes. Buoyant densities are indicated at the top of the figures (From Thiery et al., 1976).
1.3. Isochores and isochore families

The results of Thiery et al. (1976) established that gross compositional changes had occurred during the evolution of vertebrates. At the same time, they raised two problems, concerning the original size of the genome regions corresponding to the major DNA components and the physicochemical properties which were responsible for the fractionation.

(i) The problem of compositional homogeneity as related to fragment size was approached by investigating, in terms of buoyant densities and relative amounts of major
components, DNAs degraded to about 3 kb, or prepared in such a way as to obtain fragment sizes higher than 300 kb (Macaya et al., 1976). In the first case, only a modest increase in molecular heterogeneity showed up. In the second, discontinuities corresponding to the buoyant densities of the major components appeared in the CsCl profile (Fig. 3.5). These results suggested a certain degree of compositional homogeneity in the major DNA components over a very broad molecular size range, 3 kb to over 300 kb. These long, compositionally fairly homogeneous segments originating the major DNA components by degradation during DNA preparation were later called isochores for (compositionally) equal landscapes (Cuny et al., 1981).

Figure 3.5. Analytical CsCl density gradient profiles of four mouse DNA samples of different molecular weight. Molecular weights were measured from sedimentation velocity experiments except for the highest one which was estimated from the CsCl profile (From Macaya et al., 1976).

The remarkable degree of compositional homogeneity of isochores was also demonstrated by using a completely different approach, namely by analyzing the distribution in preparative density gradients of human DNA fragments (average size 50–100 kb) that carried a particular gene. The first sequence so localized was the mouse β globin gene which was detected in the 1.701 g/cm³ component, but not in the other components (Cuny et al., 1978). This was followed by the chicken ovalbumin gene found in the 1.7025 g/cm³ component (Cortadas et al., 1979). Fig. 1.10 shows that the c-mos and the c-sis oncogenes were comprised in fractions characterized by modal buoyant densities of 1.7021 and 1.7036 g/cm³ in the first case and 1.7126 g/cm³ in the second case. As shown by the c-mos example, DNA fragments containing a given, single-copy sequence may be present in more than one fraction, because of Brownian diffusion of DNA molecules in the gradient and/or because of the compositional heterogeneity of the chromosomal region where the sequence under consideration is located. Since the chromosomal segment represented by the population of overlapping fragments carrying the genes under consideration had a size of up to 100–200 kb (see Fig. 1.11), it could be concluded that base composition was fairly homogeneous over at least that size of DNA.

Other approaches provided additional information on isochores. (i) When the sequences of long genomic regions around the human α and β globin gene clusters became available,
the compositional homogeneity over long stretches of human DNA was confirmed (Aota and Ikemura, 1986). Moreover, the regions around the α and the β globin genes clusters were remarkably GC-rich and GC-poor, respectively, as already indicated by the hybridization experiments. The analysis of long sequences from data banks (Ikemura and Aota, 1988; Ikemura et al., 1990) provided examples of human isochores. Moreover, sharp isochore boundaries were demonstrated (Fukagawa et al., 1995; Stevens et al., 1999), a point to be further discussed in Part 7. (ii) Compositional mapping based on physical maps of chromosomal regions (Bernardi, 1989) provided another approach that will be described in Part 7. (iii) When the draft human genome sequences became available, constructing compositional maps over most of the human genome became possible (see Section 1.4 below).

An important step in the characterization of isochores was the isolation of the DNA components, that is of the large DNA fragments derived from them. This was first done in the case of the human and the mouse genomes (Cuny et al., 1981; Soriano et al., 1981) and later in that of the chicken genome (Olofsson and Bernardi, 1983a). The four major DNA components (the single light component had been resolved into two components, L1 and L2, by Macaya et al., 1978; the two heavy components were called H1 and H2) were prepared by two rounds of large-scale preparative centrifugation in Cs2SO4/Ag+ and characterized in several basic properties: relative amounts, GC levels, buoyant densities and compositional heterogeneities. The results obtained on the human and mouse genomes led to the following conclusions: (i) the major DNA components form at least 85% (as judged from the yields), and possibly the totality, of the main-band DNAs (Table 3.3); (ii) the relative amounts of interspersed repeats and unique sequences were strikingly different in the different major components of each genome and in the corresponding major components of the human, mouse and chicken genomes (Fig. 3.6); (iii) within each major component, the interspersed repeated sequences and the “single-copy” sequences did not significantly differ in GC levels (Fig. 3.7), another evidence for a remarkable degree of compositional homogeneity.

(ii) It is appropriate, at this point, to discuss the different features that characterized the major DNA components, some of which were responsible for their fractionation. Table 3.4 lists such differences. It is obvious that the fractionation procedure used, involving sequence-specific ligands (such as Ag+ and BAMD), took advantage of differences in the frequency of short sequences. Such frequency is expected to vary with average GC levels of DNA fragments. But clearly there is more than that. Indeed, GC values as calculated from buoyant densities (see Table 3.3) using the relationship of Schildkraut et al. (1962) which was established using bacterial DNAs, were systematically overestimated. Moreover, oligonucleotide frequencies do not vary in a monotonous way in different major components as shown by the different compositions of 3’ and 5’ termini flanking the cuts introduced by spleen DNase and snail DNase (see Fig. 3.8 and Part 1). Differences in gene density are not likely to play a role because genes generally lack common specific short-sequence features. In contrast, differences in the density and type of interspersed repeats might play a role. Minor effects might also be due to the very GC-rich CpG islands and to differences in CpG and, possibly, in methylation.
TABLE 3.3
Basic properties of mouse and human DNAs and their components

<table>
<thead>
<tr>
<th>Species</th>
<th>DNA</th>
<th>Amount</th>
<th>$\rho_0$</th>
<th>GC calculated (%)</th>
<th>H determined (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>%</td>
<td>g/cm$^3$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse</td>
<td>Main-band</td>
<td>93</td>
<td>1.700$_8$</td>
<td>41.6</td>
<td>40.6</td>
</tr>
<tr>
<td></td>
<td>1.699-g/cm$^3$</td>
<td>L1</td>
<td>22</td>
<td>1.699$_4$</td>
<td>40.2</td>
</tr>
<tr>
<td></td>
<td>1.701-g/cm$^3$</td>
<td>L2</td>
<td>34</td>
<td>1.701$_1$</td>
<td>41.9</td>
</tr>
<tr>
<td></td>
<td>1.704-g/cm$^3$</td>
<td>H1</td>
<td>26</td>
<td>1.704$_2$</td>
<td>45.1</td>
</tr>
<tr>
<td></td>
<td>1.708-g/cm$^3$</td>
<td>H2</td>
<td>7</td>
<td>1.708$_2$</td>
<td>49.2</td>
</tr>
<tr>
<td></td>
<td>Satellite</td>
<td>7</td>
<td>1.691$_5$</td>
<td>32.1</td>
<td>36.0</td>
</tr>
<tr>
<td>Human</td>
<td>Total</td>
<td>100</td>
<td>1.698$_8$</td>
<td>39.6</td>
<td>40.3</td>
</tr>
<tr>
<td></td>
<td>1.698-g/cm$^3$</td>
<td>L1</td>
<td>29</td>
<td>1.698$_1$</td>
<td>38.7</td>
</tr>
<tr>
<td></td>
<td>1.700-g/cm$^3$</td>
<td>L2</td>
<td>33</td>
<td>1.699$_9$</td>
<td>40.7</td>
</tr>
<tr>
<td></td>
<td>1.704-g/cm$^3$</td>
<td>H1</td>
<td>22</td>
<td>1.703$_8$</td>
<td>44.7</td>
</tr>
<tr>
<td></td>
<td>1.708-g/cm$^3$</td>
<td>H2</td>
<td>9</td>
<td>1.708$_1$</td>
<td>49.1</td>
</tr>
<tr>
<td>Chicken</td>
<td>Total</td>
<td>100</td>
<td>1.700$_2$</td>
<td>41.0</td>
<td>39.3</td>
</tr>
<tr>
<td></td>
<td>1.699-g/cm$^3$</td>
<td>L1</td>
<td>39</td>
<td>1.699$_0$</td>
<td>39.8</td>
</tr>
<tr>
<td></td>
<td>1.702-g/cm$^3$</td>
<td>L2</td>
<td>25</td>
<td>1.701$_9$</td>
<td>42.8</td>
</tr>
<tr>
<td></td>
<td>1.705-g/cm$^3$</td>
<td>H1</td>
<td>15</td>
<td>1.703$_6$</td>
<td>44.5</td>
</tr>
<tr>
<td></td>
<td>1.708-g/cm$^3$</td>
<td>H2</td>
<td>9</td>
<td>1.708$_4$</td>
<td>49.4</td>
</tr>
</tbody>
</table>

* (From Cuny at al. 1981). Chicken data are from Olofsson and Bernardi (1983). DNA amounts were determined from their yields. $\rho_0$ is the modal buoyant density in CsCl; the experimental error is $\pm$ 0.0005 g/cm$^3$. H, the intermolecular, or buoyant density, heterogeneity calculated as GC, was either deduced from $\rho_0$ values according to the relationship of Schildkraut et al. (1962) or determined by nucleoside analysis; the error is $\pm$ 0.5%.

TABLE 3.4
Properties of the major components of human DNA

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GC</td>
</tr>
<tr>
<td>2</td>
<td>CpG/GpC</td>
</tr>
<tr>
<td>3</td>
<td>Repeated sequences</td>
</tr>
<tr>
<td>4</td>
<td>Methylation</td>
</tr>
<tr>
<td>5</td>
<td>Short-sequence frequency</td>
</tr>
<tr>
<td>6</td>
<td>Gene density</td>
</tr>
<tr>
<td>7</td>
<td>CpG islands</td>
</tr>
</tbody>
</table>

The early hybridization experiments just mentioned were carried out on isolated major components, as described above (Cuny et al., 1981). When a human $\alpha$ globin probe failed to hybridize on any of the major components of human DNA identified until then, hybridization was done on individual Cs$_2$SO$_4$/BAMD fractions, revealing that the gene under
Figure 3.6. Reassociation kinetics of mouse and human DNAs and their major components. Results obtained with *E.coli* DNA (open circles; top left frame) are shown for the sake of comparison. The solid lines through the experimental points (solid circles) are the overall profiles resulting from the analysis of kinetic classes (broken lines). $C_{o,t}$ is the product of initial DNA concentration by reassociation time. (From Soriano et al., 1981). Chicken data are from Olofsson and Bernardi (1983).

consideration (as well as other genes tested later) in fact hybridized on fractions higher in buoyant density than the 1.708 g/cm$^3$ major component. An example of such hybridization is that for the *c-sis* oncogene shown in Fig. 1.10. This identified an additional “major” component, derived from another isochore family, which had been previously considered as a “minor component” because of its low relative amount in the human genome, and which was called H3 (Zerial et al., 1986a).

The compositional distribution of large DNA fragments from the human genome, which is a good representative of the majority of mammalian genomes (Sabour et al.,
and the most thoroughly studied, is characterized by the presence of five families of fragments derived from the corresponding isochores. In the human genome, the GC-poor isochore families L1 and L2 represent about 33% and 30% of the genome, the GC-rich H1 family corresponds to about 24% of the genome and the very GC-rich H2 and H3 families form 7.5% and 4–5% of the genome, respectively, the remaining DNA consisting of satellite and ribosomal DNAs (Zoubak et al., 1996). These percentages are slightly different from those presented in Table 3.3, which were determined on the basis of their yields from two subsequent large-scale runs of preparative centrifugation. Figs. 3.9A and 3.9B summarize the isochore organization of the human genome and its isochore pattern.

To sum up, the investigations presented in the preceding sections (Filipski et al., 1973; Thiery et al., 1976; Macaya et al., 1976) led to three major discoveries. The first one was the discovery of isochores, namely of long, compositionally fairly homogeneous regions forming most, or all, of the main bands of vertebrate genomes. For technical reasons, the average size of such fairly homogeneous regions originally could only be defined as larger than 300 kb. Incidentally, the sequences of satellite and ribosomal DNAs could also be visualized as isochores, because of their compositional homogeneity.

The second discovery was that isochores belonged to a small number of families characterized by distinct average base compositions and by other sequence features (see Table 3.4). The major DNA components were responsible for the discontinuities revealed by the CsCl analysis of Cs2SO4 ligand fractions (Fig. 3.4) and derived from isochore families that caused discontinuities in the CsCl profiles of very high molecular weight samples (Fig. 3.5).

The third discovery was that isochore patterns showed phylogenetic differences. The most striking difference concerned the genomes of warm- and cold-blooded vertebrates.
Figure 3.8. Deviation patterns of human DNA and its major components. The histogram shows the differences between the composition of the termini formed by spleen and snail DNase acting on human DNA and the compositions expected for bacterial DNAs having the same GC levels. The deviation patterns are different for total DNA and for its major components. (From Devillers-Thiéry, 1974).

This difference was the starting point of evolutionary investigations on the vertebrate genomes, to be discussed in Parts 11 and 12.

Compositional patterns were called genome phenotypes (Bernardi and Bernardi, 1986a), because they are different in different vertebrate classes (see also Part 4). It should be stressed that, until the work of Thiery et al. (1976), the common belief was that, while DNA composition varied greatly among bacteria, animal species had a rather similar DNA composition (Kit, 1960). Although some small variations in modal buoyant densities and standard deviation of CsCl profiles had been detected in vertebrate DNAs, these could be due to different methylation levels (Kit, 1962).
1.4. Isochores and the draft human genome sequence

Early in 2001, two papers reported draft sequences of the human genome. Both of them dealt with, among other subjects, the broad genomic landscape and gene density. The paper by Venter et al. (2001) summarized these features in a Table (reproduced here as Table 3.5) in which the estimates from the human sequence were compared with our estimates (Zoubak et al., 1996; Bernardi, 2000a). Venter et al. (2001) found differences between the amounts of isochores “observed” by them and those “predicted”, i.e., “based
on Bernardi's definitions of the isochore of the human genome”. These differences are, however, simply due to the fact that Venter et al. (2001) took as borders between isochore families GC levels different from ours. Indeed, the H3/H2 border was put at 48% GC instead of 50.8%, and the H1/L border at 43% instead of 41.5%. The differences in the relative amounts of genes located in different isochore families were due in part to the different estimates of isochore family borders (see above) and in part to an overestimate of genes located in GC-poor isochores (as judged by comparison with the data of Lander et al., 2001).

**TABLE 3.5**  
Characteristics of G+C in isochores

<table>
<thead>
<tr>
<th>Isochore</th>
<th>G+C (%)</th>
<th>Fraction of genome</th>
<th>Fraction of genes</th>
<th>Predicted</th>
<th>Observed</th>
<th>Predicted</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3</td>
<td>&gt;48</td>
<td>5</td>
<td>9.5</td>
<td>37</td>
<td>24.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H1/H2</td>
<td>43-48</td>
<td>25</td>
<td>21.2</td>
<td>32</td>
<td>26.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>&lt;43</td>
<td>67</td>
<td>69.2</td>
<td>31</td>
<td>48.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* From Venter et al. (2001).

The second paper (Lander et al., 2001; also referred to as International Human Genome Sequencing Consortium, 2001) studied “the draft genome sequence to see whether strict isochores could be identified” and failed to find any. They concluded that their results “rule out a strict notion of isochores as compositionally homogeneous” and that “isochores do not appear to deserve the prefix ‘iso’.”

Since the terminology “strict isochores” was misleadingly used by the authors to denote sequences that cannot be distinguished from random (uncorrelated) sequences (in which every nucleotide is free to change), their failure to identify in the human genome sequences as homogeneous as random sequences (masquerading as “strict isochores”) could have been predicted easily on three accounts.

First, for over 40 years, since at least the work of Rolfe and Meselson (1959; see also Elton, 1974), random sequences had been known to be much more homogeneous than the least heterogeneous genomic DNAs, namely bacterial DNAs (viral DNAs are not considered here because, if intact, they are perfectly homogeneous; see Fig. 1.6). In turn, bacterial DNAs are much less heterogeneous than mammalian DNAs (even if satellite and minor components are neglected).

Second, early indications that standard deviations of CsCl profiles of major components from the bovine genome were comparable to those of bacterial DNAs of the same size had already been obtained by Filipski et al. (1973). Moreover, unfraccionated DNAs from many fishes were compositionally more homogeneous than bacterial DNAs having the same fragment size and the same composition (Hudson et al., 1980; see Figs. 3.10, 3.11). Finally, Cuny et al. (1981) quantified the compositional heterogeneity of the major DNA components (namely the compositional families of 50–100 kb DNA molecules derived from isochore families; see Tables 3.1, 3.3 and Figs. 3.10, 3.11) from human and mouse.
Figure 3.10  Intermolecular compositional heterogeneity (H) plotted as a function of molecular weight (10^6 daltons is about 1500 bp) for DNAs of 33 fish species. Numbering of species is from Table 2 of Hudson et al. (1980). The solid line represents the dependence of H on fragment size for a DNA of random sequence. The broken lines show this dependence for two bacterial DNAs: E. coli (51% dG+dC, upper dashed line) and H. influenzae (38% dG+dC, lower dashed line) Data for human (○), main-band mouse (■), and chicken (●) DNAs and their components (1 to 4) are also given. Data for random DNA, bacterial, mammalian and avian DNAs are from Cuny et al., 1981. (From Hudson et al., 1980).

Figure 3.11  Comparison of heterogeneity ratios of DNAs from 33 species of fishes. Ratios were obtained by dividing intermolecular compositional heterogeneity values from each species of fish by the corresponding value of bacterial DNA having the same GC level and the same molecular weight (standard error ± 0.2) Numbering of samples is from Table 2 of Hudson et al. (1980). GC levels for fish DNAs were calculated from <ρ> values given in the same table; values given by nucleoside analysis were also used when available to calculate ratios; in this case ratios were always higher than those calculated from <ρ> and are also shown in the figure. Results of Cuny et al. (1981) for mouse main-band DNA, total human and chicken DNAs, and their components (1 to 4), including mouse satellite (5), are also shown (From Hudson et al., 1980).
and found that they were only about 30% more heterogeneous (by comparison of standard deviations) than bacterial DNAs having the same size and composition, but much more heterogeneous than random sequences. On this basis, Cuny et al. (1981) defined isochores as “fairly homogeneous regions”. It should be stressed that heterogeneity was overestimated by Cuny et al. (1981), because it was measured on pooled Cs₂SO₄/Ag⁺ fractions obtained from two subsequent large-scale centrifugations, which entailed some extent of cross-contamination between DNA components. Figs. 3.10 and 3.11 show that, in the case of chicken DNA (Cortadas et al., 1979), the heterogeneity of major components was barely different from those of reference bacterial DNAs (see above), and that unfractionated fish DNAs are, on the average, indistinguishable from bacterial DNAs, that satellite DNAs are more homogeneous than bacterial DNAs and that degradation of mouse DNA components down to 7.5 kb fragments only led to a modest increase in heterogeneity.

Third, “strict isochores” cannot exist in any natural DNA (i) because coding sequences are made up of codons, in which the compositions of the three positions are correlated with each other (D’Onofrio and Bernardi, 1992); (ii) because non-coding sequences are compositionally correlated with the coding sequences that they embed (Bernardi et al., 1985b; Clay et al., 1996; see Chapter 3 below); and (iii) because interspersed repeats are characterized by their own specific sequences. More detailed discussions of this problem were presented by Clay and Bernardi (2001a,b), Clay et al. (2001) and Clay (2001).

In summary, the conclusion of Lander et al. (2001) that “isochores” are not “strict isochores”, i.e. are not as homogeneous as random sequences is correct, but it is something we have known for at least 20 years. To take random sequences as a reference for homogeneity and looking for the same level of homogeneity in human DNA sequences was a mistake. Furthermore, it raised doubts about the very existence of isochores in readers who were not familiar with this issue, and this was unfortunate.

Around the same time, other laboratories also missed the point that “strict isochores” cannot exist in natural DNA and also took random sequences as references for compositional homogeneity (Häring and Kypr, 2001; and also, in part, Nekrutenko and Li, 2000; these papers have been commented in detail in Clay and Bernardi 2001a,b; Clay et al., 2001).

Along another line, Lander et al. (2001) did not accept the idea that the human genome is a mosaic of isochores, namely that the large-scale compositional heterogeneity is discrete or discontinuous, rather than continuously drifting, as generally believed thirty years ago and as proposed in a model by Fickett et al. (1992). This proposal was uncritically, yet enthusiastically, accepted by some authors: “From sodium chloride (sic!) centrifugation experiments, Bernardi et al. (1985) defined three major classes of genomic fragments with low, median, and high GC content, respectively and called them isochores. This description now appears to be artificial. Analyses of the complete human genome (International Human Genome Sequencing consortium, 2001) have dismissed the underlying hypothetical picture of

Figure 3.12. A colour-coded compositional map of the human chromosomes, representing 100 kb moving window plots that scan the draft human genome sequence of Lander et al. (2001). Colour codes span the spectrum of GC levels in 5 steps, from ultramarine blue (GC-poorest isochores) to scarlet red (GC-richest isochores). Grey vertical lines and bands correspond to the 5,000 gaps present in the euchromatic part of the chromosomes (Modified from Pavlick et al., 2002a.)
sharp boundaries between long homogeneous fragments: GC content turns out to vary continuously, and somewhat erratically, along chromosomes” (Galtier et al., 2001; our underlining; see below for the erratic nature of compositional variation). The authors’ rejection of the discontinuous compositional heterogeneity was due to their being apparently unaware of the detailed investigations that led to this conclusion (Filipski et al., 1973; Thiery et al., 1976; Macaya et al., 1976) and to the conclusion that isochore patterns are conserved in mammals and birds (see Part II, Chapter 2). Moreover, Lander et al. (2001) seem to have overlooked the evidence that GC-rich isochores arose at the transition between cold- and warm-blooded vertebrates from regions that were much lower in GC level and, in fact, very similar to the GC-poor isochores of the human genome (see Section 1.5 below and Part II). Obviously, this emergence from the more homogeneous compositional spectrum of cold-blooded vertebrates was the primary source for a discontinuous distribution (see Fig. 12.5).

Figure 3.13. A colour-coded compositional map of human chromosomes 1-6. Other indications as in Fig 3.12. (From Pavlícek et al., 2002a.)
Figure 3.14. A colour-coded compositional map of human chromosomes 7-12. Other indications as in Fig 3.12. (From Pavliček et al., 2002a.)

The denial of the very existence of isochores, if correct, would have major consequences which, apparently, were not realized by Lander et al. (2001). The first one would be the denial of a compositionally discontinuous sequence organization and the return to the continuous compositional spectrum for the human genome (see above) that was the predominant view until the work of Filipski et al. (1973). The second consequence would be the denial of “an important level of genome organization, insofar as gene density (Zoubak et al., 1996), gene length (Duret et al., 1995), and patterns of codon usage (Sharp et al. 1995), as well as the distribution of different classes of repetitive elements (Soriano et al., 1983, Duret et al., 1995), are all correlated with GC content” (Fullerton et al., 2001). Other properties that could be added to the list are replication timing, recombination frequency (Fullerton et al., 2001), chromosomal banding, and stability and transcription of integrated sequences. In other words, the second consequence would be the denial of “a fundamental level of genome organization” (Eyre-Walker and Hurst, 2001).

Very interestingly, the best graphical display of the mosaic organization of the human
Figure 3.15 A colour-coded compositional map of human chromosomes 13-18. Other indications as in Fig. 3.12. (From Pavliček et al., 2002a.)

Other indications as in Fig. 3.12. (From Pavliček et al., 2002a.)

The genome available at present is that of the compact compositional maps of human chromosomes (Figs. 3.12-3.16) derived from the data of Lander et al. (2001) which confirm, a quarter of a century later, the isochore structure of mammalian genomes (Macaya et al., 1976). Indeed, apart from the presence of about 5,000 gaps (grey bars) in the euchromatic regions of most chromosomes (in fact of all chromosomes, except chromosomes 21 and 22, which had been sequenced previously by two other groups, Hattori et al., 2000, and Dunham et al., 1999), the striking feature of the map is undoubtedly the large proportion of the genome represented by long GC-poor regions, uninterrupted by GC-rich regions. The next most notable observation is the scarcity of GC-poor regions in many of the blocks characterized by GC-rich regions. As far as these two points are concerned, the results fit with the previous estimates of the relative amounts of GC-poor and GC-rich isochores (see Fig. 3.8) and with the very high yields of major DNA components from the human genome (Cuny et al., 1981). They contradict the suggestion (Eyre-Walker and Hurst, 2001) that an isochore structure accounts for “only some parts” of the genome. The third observation is the increasing compositional fluctuations when moving from GC-
poor to GC-rich isochores. This had already been noticed in previous work (Cuny et al., 1981; De Sario et al., 1996; 1997) and was confirmed by a detailed analysis of chromosomes 21 and 22 (Saccone et al., 2001). A working hypothesis on this point is presented in Part 7, Chapter 3.

Finally, it should be noted that the considerations presented in this section were not affected by the recent advances towards completing the human genome sequence (see, for example, Heilig et al., 2003; Mungall et al., 2003).

1.5. Other misunderstandings about isochores

Another article (Nekrutenko and Li, 2000) also exemplifies some common misunderstandings about isochores, but, in addition, proposes new criteria and measures of heterogeneity that are relevant to the results presented here. Indeed, the authors propose a new hetero-
geneity parameter (apparently as an alternative to the standard deviation), namely the average magnitude of the difference between GC levels of successive segments of length \( l \) in an \( n \)-segment partitioning (or covering by partially overlapping windows) of a DNA sequence of interest (after appropriate normalization):

\[
\Delta GC := \frac{1}{n-1} \sum_{i=2}^{n} |GC_i - GC_{i-1}|
\]

In fact, the quantity \( \Delta GC \), independently used by us in a comparative analysis of the human and \textit{Drosophila} genome (Jabbari and Bernardi, 2000), has a nice graphical interpretation. If one plots the GC level of each segment \( i \) against the GC level of the preceding segment \( i-1 \), one obtains, in human and even in \textit{Drosophila} DNA, a good correlation, with the points of the scatter plot lining up near the main diagonal of slope 1 (see \textbf{Part 9}; \textbf{Fig. 9.7}). Such plots are called “time-delay plots” in physics (Ruelle, 1989), or “phase plots”, since they can show the phase of a periodic fluctuation. They can be used to detect and display correlations along DNA sequences, if nonoverlapping windows are chosen. It can be easily seen that \( \Delta GC \) is simply the mean vertical (or horizontal) distance of the points from the main diagonal; the greater the (absolute) GC differences between successive segments, the farther the points from the diagonal, and the higher the heterogeneity.

It is, therefore, natural that Nekrutenko and Li (2000) also suggest a criterion, based on the differences \( |GC_i - GC_{i-1}| \), for the detection of isochore boundaries. Clearly, the points representing pairs of adjacent segments within a single GC-poor isochore will cluster around one part of the diagonal, those representing an adjacent GC-rich isochore will cluster around another part of the diagonal, and the few segment pairs spanning a sharp boundary between two adjacent isochores will be found far from the diagonal. A natural question is how far the points need to be from the diagonal, i.e., how big \( |GC_i - GC_{i-1}| \) needs to be, before it is indicative of an isochore boundary. Any such critical distance must depend on the variation in GC level that can be expected within a single isochore (i.e., depend on the fragments’ GC distribution), and on some chosen significance level such as 5%, 1% or 0.1%.

Unfortunately, the authors did not resolve this problem. Their first normalization is incorrect: not only did they choose a binomial distribution, \textit{i.e.}, an uncorrelated DNA sequence, as the expected value for normalizing (which, as argued above, and as subsequently realized also by the authors, is inappropriate), but they neglected that the absolute difference \( |GC_i - GC_{i-1}| \) has a standard deviation that is \( 2/\sqrt{\pi} \approx 1.128 \) times larger than that of \( GC_n \), \textit{i.e.}, than that of a simple binomial distribution or of its normal approximation. A factor \( \sqrt{2/\pi} \) arises because the variance of a distribution of differences of two random variables is, in the correlation-free case considered here, the sum of the variances of their distributions; an additional factor \( \sqrt{2/\pi} \) is introduced by taking absolute values (and using the approximate normality of the binomial distributions of interest; see Shiryayev, 1984, pp. 237 ff., or Bourbaki, 1969, p. 74). As a result, the authors' 400,000 simulations for different fragment lengths and overlaps (Nekrutenko and Li, 2000, Table 1), using randomly generated, uncorrelated sequences, repeatedly give values close to 1.12 – 1.13, namely \( 2/\sqrt{\pi} \), instead of 1 as expected. This apparent “behavior” or “baseline value” of their compositional heterogeneity index is, therefore, simply an artefact of incorrect normalization. In addition, since the authors use the binomial distribution, \textit{i.e.}, sequences of
independent nucleotides, as their first standard for homogeneity, they find no long homogeneous regions and are obliged to search for a more appropriate standard. (A similar two-step reasoning, amounting to a rediscovery of the results of Cuny et al., 1981, can be discerned also in Lander et al., 2001, and in Häring and Kypr, 2001, a study that has been discussed in more detail by Clay and Bernardi, 2001a.)

The second attempt by Nekrutenko and Li (2000) to compare $|GC_i - GC_{i+1}|$ with a control distribution, namely by assuming that intra-isochrome fluctuations in human should not exceed typical intra-chromosomal fluctuations in yeast, the smallest ones observed so far in an eukaryotic genome (see Part 9), is no less arbitrary and extrinsic, but it does lead to somewhat longer “homogeneous” regions than when random sequences are used as a yardstick. This partial success leads the authors rather quickly to a bold proposal, namely to a “new definition of an isochore as any genomic fragment longer or equal to 100 kb such that when it is divided into a series of overlapping 10-kb windows, no two windows can differ by $>7\%$ GC”. In addition to the underestimate of intra-isochrome fluctuations imposed by their yeast threshold, there is a second problem inherent in their approach: the probability of witnessing at least one unlikely event (such as 4 consecutive heads when tossing a coin) will always increase with the number of trials, so that working outwards from an “isochore seed” until a given fluctuation limit is reached will make it increasingly difficult for long isochromes to meet the authors’ criterion. It does not come as a surprise, therefore, that their method results in many very short homogeneous fragments, and in no long ones. The short ones are duly eliminated by the artificially imposed lower bound of 100 kb, and banished to regions of DNA purported to contain no isochromes at all, while long “isochromes” remain conspicuously absent. For example, in chromosome 21, where most moving window plots immediately reveal a strikingly homogeneous, long, very GC-poor region covering more than 7 Mb (De Sario et al., 1997; Hattori et al., 2000; Oliver et al., 2001; Li, 2001; see Fig. 3.16), the authors’ histogram shows only two “isochromes” longer than 350–370 kb (with lengths of $\approx 430$ and $\approx 490$ kb, respectively). Clearly, such a redefinition of isochromes cannot be considered tenable, since it could furnish, at best, only the shortest isochromes in the human genome.

Recent results show that isochromes and their boundaries can be recognized not only by color-coded moving window plots (Pavliček et al., 2002a; see Figs. 3.12–3.16), but also by entropic segmentation (Oliver et al., 2001; Li, 2001), and by simple windowless method (Zhang and Zhang, 2003a,b). Interestingly, this approach was applied by Macaya and Bernardi thirty years ago, but its use on total DNA fragments about 50 kb in size was not (and could not be) as successful as when applied to DNA sequences. Statistical correlations in DNA sequences have been investigated by Bernaola-Galvan et al. (2002) and isochore chromosome maps of the human genome have been published by Oliver et al. (2002). The principle and the applications are described in Part 9, Chapter 5.3. Furthermore, if traditional statistical procedures are correctly applied, they reveal the existence of a significant relative homogeneity in GC within isochromes (Li et al., 2003; this paper critically discusses the calculations of Lander et al., 2001). Interestingly, the isochore structure of the mammalian genome has reached the sequencing community, as witnessed by the number of compositional plots presented by the Mouse Genome Sequencing Consortium (2002).
Compositional patterns of coding sequences

Fig. 3.17A presents a schematic view of the isochore patterns of the vertebrate genomes best studied by density gradient centrifugation. The relative amounts, modal buoyant densities and GC levels of isochore families are shown. Apart from the major differences between the *Xenopus* and human patterns, the latter showing GC-rich isochore families that are absent in *Xenopus*, it is remarkable that the human pattern is also slightly yet significantly different from the chicken pattern, which exhibits an additional very GC-rich family H4, and from the mouse pattern, which lacks the GC-rich family H3. These differences will be discussed in more detail in Part 4.

Very remarkably, the differences in isochore patterns of vertebrate genomes were paralleled by differences in base composition (Fig. 3.17B) of coding sequences (Bernardi et al., 1985b, 1988; Perrin and Bernardi, 1987; Bernardi and Bernardi, 1991) which only represent a very small minority of the genomes under consideration. A first comparison could be made by looking at the compositional distribution of coding sequences, or of their average third codon positions, GC3. This showed striking differences not only between human and *Xenopus*, but also between human and chicken, and between human and mouse. Indeed, the compositional distribution of coding sequences from *Xenopus* is narrower and centered on a lower value compared to the human distribution. The distribution of chicken genes, also seen as GC3 values, reached a higher level, in fact 100%, compared to the human distribution. Finally, differences could also be seen between human and mouse, the latter distribution being narrower.

Figure 3.17. A. Isochore families from *Xenopus*, human, mouse and chicken, as deduced from density gradient centrifugation. B. Compositional patterns of coding sequences (represented by GC3 values averaged per coding sequence) for *Xenopus*, mouse, chicken and human. (Modified from Bernardi, 1995).
CHAPTER 3

Compositional correlations between coding and non-coding sequences

The similar compositional features exhibited by isochore families and by coding sequences of the genomes of Fig. 3.17 raised the question of the possible correlation between the composition of the isochores, which cover a 30 to 60% GC range in the case of the human genome, and the composition of the coding sequences embedded in them. This question could be answered by localizing coding sequences of known primary structure in compositional DNA fractions (or in isolated major DNA components).

As already mentioned, hybridization of appropriate probes on compositional fractions of genomic DNA (Fig. 1.10) allows localizing sequences in such fractions, defining the GC levels of those fractions and exploring chromosomal regions up to twice the size of the DNA fragments making up the DNA samples (see Fig. 1.11). When a number of coding sequences, essentially from human, mouse and chicken, were localized by hybridization of appropriate probes, and their GC levels were plotted against GC levels of the major components or compositional fractions in which they had been localized, linear correlations were found (Bernardi et al., 1985a,b; preliminary reports were published by Bernardi, 1979b; 1984). Linear correlations were also found when plotting GC values for entire genes (exons + introns; Fig. 3.18) and for introns, or GC3 values, against those of the corresponding isochores. These results were of great relevance because they showed compositional correlations between the coding sequences (that represent 2–3% of the genome) and the non-coding sequences (introns and intergenic sequences, that represent 97–98% of the genome), in which, or next to which, they were located. It should be noted that a correlation between GC levels in third positions and GC levels of flanking sequences was independently reported by Ikemura (1985) for some vertebrate genes.

When GC and nucleotide levels of the three codon positions of prokaryotic, viral and vertebrate genes were plotted against GC levels of the corresponding genome or (in the case of vertebrate genes) of exons, linear correlations with high correlation coefficients were found in all cases (Bernardi and Bernardi, 1985; 1986a,b; Fig. 3.19). Incidentally, the results on prokaryotic GC correlations (top frames of Fig. 3.19) were later confirmed on a small gene sample by Muto and Osawa (1987).

The localization of coding sequences was further studied when long genomic sequences from human DNA became available. In this case, direct comparisons could be done between the GC levels of coding and non-coding sequences and more precise estimates of the correlations could be obtained (Aïssani et al., 1991; Clay et al., 1996; Zoubak et al., 1996). Figure 3.20A shows that linear correlations hold between the GC3 levels of coding sequences and the GC levels of the isochores in which coding sequences are located. Interestingly, GC3 values of GC-poor coding sequences and their flanking sequences show very similar values, whereas GC3 values of GC-rich coding sequences are increasingly higher above the diagonal, essentially because GC3 values depart more and more
from the intergenic sequences. Linear correlations also hold (Fig. 3.20B) between the GC levels of coding sequences and the GC levels of the introns of the same genes (Bernardi et al., 1985b; Aissani et al., 1991; Clay et al., 1996). The values of introns were systematically lower than those of exons, a point commented upon in Part 12, Chapter 2.

Figure 3.18. Plot of GC content of genes against the GC levels of DNA components in which they are located. The numbers indicate genes (see the original article). The line was drawn using the least-square method. The unit slope line corresponds to the coincidence in GC contents of genes and major components in which genes are located. (Modified from Bernardi et al., 1985b).
Figure 3.19. GC and nucleotide levels of the three codon positions (1, 2, 3) of prokaryotic, viral and vertebrate genes plotted against GC levels of the corresponding genomes (or exons in the case of vertebrates; filled circles are average GC₃ levels of genes belonging to the same compartment of a given genome).

For details about the genes studied, see the original paper. (From Bernardi and Bernardi, 1986a).
Figure 3.20. A. Correlation between $\text{GC}_3$ of coding regions of genes and the GC level of fractions or Yeast Artificial Chromosomes (YACs) in which the genes were localized (filled circles), and of 3' flanking sequences farther than 500 bp from the stop codon (open circles). The corresponding, essentially coincident orthogonal regression lines, indicated here by a single dashed line, are given by $\text{GC}_3 = 3.45 \times \text{GC} - 94.6$ ($R = 0.82$, $N = 32$; fractions and YACs) and $\text{GC}_3 = 3.39 \times \text{GC} - 88.2$ ($R = 0.56$, $N = 103$; far 3' sequences). The relation obtained by comparing the independent Gaussian decomposition of GC and $\text{GC}_3$ distribution is indicated by a solid line, and given by $\text{GC}_3 = 2.92 \times \text{GC} - 74.3$. This solid line happens to be almost indistinguishable from the regression line for those genes that were known to contain no CpG islands, $\text{GC}_3 = 2.89 \times \text{GC} - 71.9$ ($R = 0.87$, $N = 13$; fractions and YACs) and $\text{GC}_3 = 2.90 \times \text{GC} - 66.4$ ($R = 0.67$, $N = 30$; far 3' sequences). (From Zoubak et al., 1996). B. Correlation between GC levels of human coding sequences and the GC levels of the corresponding introns. The diagonal (unity slope line) is also shown (Modified from Clay et al., 1996).

We have already reported that the use of the orthogonal regression as done in Fig. 3.20A, i.e., regression along the principal eigenvector of the variance-covariance matrix, gives a better description of the scatter diagrams than linear regression, since the orthogonal regression line is obtained by minimizing the sum of squares of the orthogonal distances parallel to the ordinate (see D’Onofrio et al., 1991; Mouchiroud et al., 1991). Indeed, as stressed by Cruveiller et al. (2003), bivariate distributions often have an approximately linear shape and are well characterised by their major axis (also called orthogonal regression line, or principal axis). In passing, it should be mentioned that traditional linear regression lines (such as are used to described unilateral dependence relationships) do not provide a satisfactory characterisation, since they do not follow the points when scatterplots are characterised by step slopes, but instead systematically slice the scatterplots at a lower angle (Jolicoeur, 1990; Harvey and Pagel, 1991; Mouchiroud et al., 1991; Clay et al., 1996).

As a final remark, one could say that the compositional correlations between coding and non-coding sequences and between the three codon positions (the genomic code) have shown, to paraphrase Galileo, that the book of the genome is written in a mathematical language.
Part 4

The compositional patterns of vertebrate genomes
CHAPTER 1

The fish genomes

1.1. Compositional properties: a CsCl analysis

There were several reasons to make the vertebrate genome the central subject of our investigations on structural and evolutionary genomics: the vast knowledge available on the biology of vertebrates, the existence of a good paleontological record, and the fact that vertebrates are a small taxon, which share many basic properties. But, obviously, the decisive factor was the discovery of massive compositional changes at the transitions between cold- and warm-blooded vertebrates (Thiery et al., 1976). On the one hand, this raised important questions as to how this could have occurred and which were the causes; on the other it provided an experimental access to its analysis.

As far as fishes are concerned, the work was done in four steps involving a total of 2 (Thiery et al., 1976), 34 (Hudson et al., 1980), 39 (Pizon, 1984), 122 (Bernardi and Bernardi, 1990a) and 201 (Bucciarelli et al., 2002) species, respectively. The reasons for exploring such a large number of fish genomes was (i) to obtain a representative sample of different orders; fishes include a vast array of distantly related vertebrates with about 25,000 species, corresponding to about half of the extant vertebrate species (Nelson, 1994); (ii) to extend as much as possible comparisons of fish orders, families, and genera, in order to investigate the compositional differences that arose over evolutionary time (see Bernardi and Bernardi, 1990b); and (iii) to draw some general conclusions about the organization of fish genomes.

Table 4.1 presents a classification of fishes derived from Nelson (1994) in order to indicate the taxonomic position of the species studied. This classification puts the fishes in an order that reflects their postulated evolutionary relationship, ranging from the most ancient splits in vertebrate evolution, Chondrichthyes (sharks, rays), Sarcopterygii (Lungfish, Coelacanths and Tetrapods) and Actinopterygii (ray-finned fishes), to the most recent teleost orders, Pleuronectiformes (flounders) and Tetraodontiformes (puffers).

Our study included representatives from 3 out of 9 orders of Elasmobranchi (sharks and rays), both orders of dipnoan lungfishes, and both orders of chondrosteans (sturgeons and bichirs). We also studied 19 out of 38 teleostean orders, which represent all but 4 (minor) superorders of the subdivision Teleostei, a group comprising about 23,600 species (96% of all extant fishes). This leaves for further studies two subclasses, Holocephali (chimaeras), and Coelacanthimorpha (gombessas). Table 4.2 lists all species studied with their properties.

Figure 4.1 displays the analytical CsCl profiles of the fish DNAs studied by Bernardi and Bernardi (1990a). Only three DNA samples (4, 11, 12; all of them from Chondrichthyes) showed sizable amounts of resolved satellite bands, which were on the heavy, GC-rich side of the main band. Minor satellite bands were found in other DNA samples, mainly on the
<table>
<thead>
<tr>
<th>Class</th>
<th>Subclass</th>
<th>Infraclss</th>
<th>Division</th>
<th>Subdivision</th>
<th>Superorder</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chondrichthyes</td>
<td>*Holocephali</td>
<td></td>
<td></td>
<td></td>
<td>Euselachi</td>
<td>Lamniformes</td>
</tr>
<tr>
<td></td>
<td>Elasmobranchi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Squaliformes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rajiformes</td>
</tr>
<tr>
<td>Sarcopterygii</td>
<td>*Coelacanthimorpha</td>
<td></td>
<td></td>
<td>Ceratodontmorpha</td>
<td>Ceratodontiformes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dipnoi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Leptosireniformes</td>
</tr>
<tr>
<td></td>
<td>(Tetrapoda)</td>
<td></td>
<td></td>
<td></td>
<td>Polypiertormorphes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chondrostei</td>
<td></td>
<td></td>
<td>Ceratodontmorpha</td>
<td>Ceratodontiformes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lepidosireniformes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Neopterygii</td>
<td></td>
<td></td>
<td>Teleostei</td>
<td>Osteoglossomorpha</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Elopomorpha</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Clupeomorpha</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Euteleostei</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Protacanthopterygii</td>
<td></td>
<td></td>
<td>Ostariophysi</td>
<td>Protacanthopterygii</td>
<td></td>
</tr>
<tr>
<td></td>
<td>*Stenopterygii</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cyclostylegata</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>*Scopelomorpha</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>*Lampriomorpha</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>*Polymixiomorpha</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class</td>
<td>Subclass</td>
<td>Infraclasse</td>
<td>Division</td>
<td>Subdivision&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Superorder</td>
<td>Order&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>---------------</td>
<td>----------</td>
<td>-------------</td>
<td>----------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>----------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Paracanthopterygii</td>
<td>Gadiformes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ophidiformes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Batrachoidiformes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lophiiformes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Acanthopterygii</td>
<td>Cyprinodontiformes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Syngnathiformes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Dactylopteriformes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Scorpaeniformes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Perciformes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pleuronectiformes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Tetraodontiformes</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> From Nelson, 1994.

<sup>b</sup> Asterisks indicate groups not investigated.

<sup>c</sup> Only orders listed in Table 4.2 are presented.
### TABLE 4.2
Properties of DNAs from fishes

<table>
<thead>
<tr>
<th>Sample Order number</th>
<th>Family</th>
<th>Species</th>
<th>MW  (kb)</th>
<th>( \rho_0 ) (g/cm³)</th>
<th>(&lt;\rho&gt;) (g/cm³)</th>
<th>A (mg/cm³)</th>
<th>H %GC</th>
<th>GC %</th>
<th>e pg</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Latimeriformes</td>
<td>Scyllorhinidae</td>
<td>Scyllorhinus stellaris</td>
<td>29.55</td>
<td>1.7053</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Carchariformes</td>
<td>Mustelus ater</td>
<td></td>
<td>34.55</td>
<td>1.7044</td>
<td>1.7047</td>
<td>0.3</td>
<td>2.9</td>
<td>45.6</td>
</tr>
<tr>
<td>3</td>
<td>Sclerodactylidae</td>
<td>Sclerodactylus erythrophthalmus</td>
<td></td>
<td>13.98</td>
<td>1.7030</td>
<td>1.7035</td>
<td>0.5</td>
<td>2.9</td>
<td>44.4</td>
</tr>
<tr>
<td>4</td>
<td>Carcharhinidae</td>
<td>Carcharodon veranyas</td>
<td></td>
<td>15.24</td>
<td>1.7033</td>
<td>1.7037</td>
<td>0.4</td>
<td>2.3</td>
<td>44.7</td>
</tr>
<tr>
<td>5</td>
<td>Squalidae</td>
<td>Sphecus latus</td>
<td></td>
<td>1.57</td>
<td>1.7030</td>
<td>1.7042</td>
<td>1.2</td>
<td>4.6</td>
<td>45.1</td>
</tr>
<tr>
<td>6</td>
<td>Squalidae</td>
<td>Squale ater</td>
<td></td>
<td>48.50</td>
<td>1.7050</td>
<td>1.7059</td>
<td>0.9</td>
<td>4.3</td>
<td>46.8</td>
</tr>
<tr>
<td>7</td>
<td>Squatinidae</td>
<td>Squatinus ater</td>
<td></td>
<td>29.55</td>
<td>1.7050</td>
<td>1.7072</td>
<td>2.2</td>
<td>5.0</td>
<td>48.2</td>
</tr>
<tr>
<td>8</td>
<td>Rajiformes</td>
<td>Rajidae</td>
<td>Rajia</td>
<td>33.25</td>
<td>1.7035</td>
<td>1.7041</td>
<td>0.6</td>
<td>3.7</td>
<td>45.0</td>
</tr>
<tr>
<td>9</td>
<td>Rajia</td>
<td>Rajia ater</td>
<td></td>
<td>11.70</td>
<td>1.7014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Dasyatidae</td>
<td>Gymnura ater</td>
<td></td>
<td>44.44</td>
<td>1.7015</td>
<td>1.7022</td>
<td>0.7</td>
<td>2.7</td>
<td>43.1</td>
</tr>
<tr>
<td>11</td>
<td>Myliobatidae</td>
<td>Myliobatis frenivallaci</td>
<td></td>
<td>33.25</td>
<td>1.7029</td>
<td>1.7039</td>
<td>1.0</td>
<td>4.4</td>
<td>44.8</td>
</tr>
<tr>
<td>12</td>
<td>Torpedinidae</td>
<td>Torpedo marmorata</td>
<td></td>
<td>45.36</td>
<td>1.7011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Torpedo</td>
<td>Torpedo ater</td>
<td></td>
<td>49.46</td>
<td>1.7012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Ceratodontiformes</td>
<td>Ceratodontidae</td>
<td>Neceratodus forsteri</td>
<td>1.7031</td>
<td>1.7027</td>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Lepidodermiformes</td>
<td>Protorhineida</td>
<td>Protorhineida sp.</td>
<td>55.15</td>
<td>1.7005</td>
<td>1.7001</td>
<td>0.4</td>
<td>2.5</td>
<td>40.9</td>
</tr>
<tr>
<td>16</td>
<td>Polypteriformes</td>
<td>Polypterida</td>
<td>Polypterus senegalus</td>
<td>28.84</td>
<td>1.6995</td>
<td>1.7008</td>
<td>1.3</td>
<td>2.3</td>
<td>41.6</td>
</tr>
<tr>
<td>17</td>
<td>Acipenseriformes</td>
<td>Acipenserida</td>
<td>Acipenser steve</td>
<td>1.7019</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Osteoglossiformes</td>
<td>Osteoglossida</td>
<td>Osteoglossus myleus</td>
<td>5.53</td>
<td>1.7022</td>
<td>1.7023</td>
<td>0.1</td>
<td>2.0</td>
<td>43.2</td>
</tr>
<tr>
<td>19</td>
<td>Notopterida</td>
<td>Notopterida</td>
<td>Notopterus notopterus</td>
<td>27.91</td>
<td>1.7015</td>
<td>1.7021</td>
<td>0.6</td>
<td>3.6</td>
<td>42.9</td>
</tr>
<tr>
<td>20</td>
<td>Mormyridae</td>
<td>Mormyrida</td>
<td>Mormyrida petersii</td>
<td>24.58</td>
<td>1.7029</td>
<td>1.7033</td>
<td>0.4</td>
<td>2.5</td>
<td>44.2</td>
</tr>
<tr>
<td>21</td>
<td>Anguilliformes</td>
<td>Anguillida</td>
<td>Anguilla anguilla</td>
<td>13.68</td>
<td>1.7018</td>
<td>1.7031</td>
<td>1.3</td>
<td>2.3</td>
<td>44.0</td>
</tr>
<tr>
<td>22</td>
<td>Anguilla</td>
<td>Anguilla rostrata</td>
<td></td>
<td>35.07</td>
<td>1.7011</td>
<td>1.7015</td>
<td>0.7</td>
<td>3.5</td>
<td>42.6</td>
</tr>
<tr>
<td>23</td>
<td>Chameleiformes</td>
<td>Chamaeleida</td>
<td>Chamaeleon tiryanus</td>
<td>20.52</td>
<td>1.7021</td>
<td>1.7024</td>
<td>0.3</td>
<td>2.5</td>
<td>43.4</td>
</tr>
<tr>
<td>24</td>
<td>Cypriniformes</td>
<td>Cyprinida</td>
<td>Cyprinus auratus</td>
<td>20.14</td>
<td>1.7022</td>
<td>1.7022</td>
<td>0.0</td>
<td>3.3</td>
<td>43.1</td>
</tr>
<tr>
<td>25</td>
<td>Cyprinida</td>
<td>Cyprinida</td>
<td>Cyprinus carpio</td>
<td>24.36</td>
<td>1.6970</td>
<td>1.6971</td>
<td>0.1</td>
<td>1.8</td>
<td>37.9</td>
</tr>
<tr>
<td>26</td>
<td>Trachinidae</td>
<td>Trachinida</td>
<td>Trachinida ater</td>
<td>72.03</td>
<td>1.6963</td>
<td>1.6965</td>
<td>0.2</td>
<td>2.0</td>
<td>37.2</td>
</tr>
<tr>
<td>27</td>
<td>Brachydeipida</td>
<td>Brachydeipida</td>
<td>Brachydeipida ater</td>
<td>49.46</td>
<td>1.6959</td>
<td>1.6962</td>
<td>0.3</td>
<td>2.6</td>
<td>36.9</td>
</tr>
<tr>
<td>28</td>
<td>Labecidae</td>
<td>Labecidae</td>
<td>Labecida bicolor</td>
<td>18.65</td>
<td>1.6959</td>
<td>1.6965</td>
<td>0.7</td>
<td>2.0</td>
<td>37.2</td>
</tr>
<tr>
<td>29</td>
<td>Cobitidae</td>
<td>Cobitidae</td>
<td>Cobitis ater</td>
<td>1.6975</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TABLE 4.2
Properties of DNAs from fishes* (continued)

<table>
<thead>
<tr>
<th>Sample Order number</th>
<th>Family</th>
<th>Species</th>
<th>MW (kb)</th>
<th>(\rho_0) (g/cm³)</th>
<th>(&lt;\rho&gt;) (g/cm³)</th>
<th>A (mg/cm³)</th>
<th>H</th>
<th>%GC</th>
<th>GC %</th>
<th>e (pg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>Characiformes</td>
<td>Characidae</td>
<td>Astyanax mexicanus</td>
<td>1.6995</td>
<td>1.6998</td>
<td>0.3</td>
<td>40.6</td>
<td>44.4</td>
<td>1.1–2.1</td>
<td>38.8</td>
</tr>
<tr>
<td>31</td>
<td>Siluriformes</td>
<td>Callichthyidae</td>
<td>Corydoras aeneus</td>
<td>1.6979</td>
<td>1.6980</td>
<td>0.8</td>
<td>44.4</td>
<td>44.7</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Salmoniformes</td>
<td>Salmonidae</td>
<td>Salmo salar</td>
<td>1.7028</td>
<td>1.7035</td>
<td>0.7</td>
<td>44.4</td>
<td>44.7</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Salmoniformes</td>
<td>Salmonidae</td>
<td>Salmo fario</td>
<td>1.7030</td>
<td>1.7039</td>
<td>0.9</td>
<td>44.4</td>
<td>44.7</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
<td>Onchorhynchus keta</td>
<td></td>
<td>30.51</td>
<td>1.7038</td>
<td></td>
<td>44.4</td>
<td>44.7</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
<td>Onchorhynchus kisutch</td>
<td></td>
<td>31.24</td>
<td>1.7036</td>
<td></td>
<td>44.4</td>
<td>44.7</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td></td>
<td>Onchorhynchus mykiss</td>
<td></td>
<td>28.14</td>
<td>1.7024</td>
<td></td>
<td>44.4</td>
<td>44.7</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td></td>
<td>Onchorhynchus nerka</td>
<td></td>
<td>27.68</td>
<td>1.7035</td>
<td></td>
<td>44.4</td>
<td>44.7</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td></td>
<td>Coregonus autumnalis migr.</td>
<td></td>
<td>3.49</td>
<td>1.7034</td>
<td></td>
<td>44.4</td>
<td>44.7</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>Aulopiformes</td>
<td>Synodontidae</td>
<td>Synodus foetens</td>
<td>32.24</td>
<td>1.7025</td>
<td></td>
<td>43.8</td>
<td>43.6</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td>Synodus intermedius</td>
<td></td>
<td>1.31</td>
<td>1.7021</td>
<td></td>
<td>43.8</td>
<td>43.6</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td></td>
<td>Trachinocephalus myops</td>
<td></td>
<td>31.00</td>
<td>1.7034</td>
<td></td>
<td>43.8</td>
<td>43.6</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>Gadiformes</td>
<td>Gadidae</td>
<td>Urophycis chuss</td>
<td>5.86</td>
<td>1.7070</td>
<td></td>
<td>47.2</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td></td>
<td>Urophycis regius</td>
<td></td>
<td>22.29</td>
<td>1.7065</td>
<td></td>
<td>47.2</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td></td>
<td>Merluccius bilinearis</td>
<td></td>
<td>52.43</td>
<td>1.7041</td>
<td></td>
<td>47.2</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>Ophidiformes</td>
<td>Ophidiidae</td>
<td>Ophidion holbrooki</td>
<td>6.43</td>
<td>1.7018</td>
<td></td>
<td>43.6</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>Batrachoidiformes</td>
<td>Batrachoididae</td>
<td>Opsanus tau</td>
<td>23.73</td>
<td>1.6996</td>
<td></td>
<td>40.9</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td></td>
<td>Porichthys notatus</td>
<td></td>
<td>46.00</td>
<td>1.6979</td>
<td></td>
<td>40.9</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td></td>
<td>Porichthys porosissimus</td>
<td></td>
<td>42.34</td>
<td>1.6998</td>
<td></td>
<td>40.9</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>Lophiiformes</td>
<td>Lophidae</td>
<td>Lophius americanus</td>
<td>13.08</td>
<td>1.6996</td>
<td></td>
<td>40.9</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>Cyprinodontiformes</td>
<td>Aplocheilidae</td>
<td>Aplocheilus days</td>
<td>1.7002</td>
<td>1.7008</td>
<td></td>
<td>41.6</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td></td>
<td>Aphyosemion amieti</td>
<td></td>
<td>13.38</td>
<td>1.7016</td>
<td></td>
<td>41.6</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td></td>
<td>Aphyosemion australis</td>
<td></td>
<td>65.34</td>
<td>1.7066</td>
<td></td>
<td>41.6</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td></td>
<td>Aphyosemion cameronense</td>
<td></td>
<td>15.40</td>
<td>1.7009</td>
<td></td>
<td>41.6</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td></td>
<td>Aphyosemion elegans</td>
<td></td>
<td>13.23</td>
<td>1.7008</td>
<td></td>
<td>41.6</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td></td>
<td>Aphyosemion hertzogii</td>
<td></td>
<td>36.95</td>
<td>1.7013</td>
<td></td>
<td>41.6</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td></td>
<td>Aphyosemion marmaratum</td>
<td></td>
<td>25.22</td>
<td>1.7018</td>
<td></td>
<td>41.6</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>57</td>
<td></td>
<td>Aphyosemion punctatum</td>
<td></td>
<td>6.43</td>
<td>1.7009</td>
<td></td>
<td>41.6</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td></td>
<td>Aphyosemion schelii</td>
<td></td>
<td>17.59</td>
<td>1.7011</td>
<td></td>
<td>41.6</td>
<td>45.4</td>
<td>0.9</td>
<td></td>
</tr>
</tbody>
</table>
TABLE 4.2
Properties of DNAs from fishes* (continued)

<table>
<thead>
<tr>
<th>Sample Order number</th>
<th>Family</th>
<th>Species</th>
<th>MW (kb)</th>
<th>( \rho_0 ) (g/cm(^3))</th>
<th>( &lt;\rho&gt; ) (g/cm(^3))</th>
<th>A (mg/cm(^3))</th>
<th>H %GC</th>
<th>GC %</th>
<th>c (pg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>59</td>
<td>Aphyoseminon</td>
<td>Aphyoseminon spoenoerbergi</td>
<td>10.60</td>
<td>1.7032</td>
<td>1.7032</td>
<td>0.0</td>
<td>1.9</td>
<td>44.1</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>Aphyoseminon</td>
<td>Aphyoseminon striatum</td>
<td>10.47</td>
<td>1.7053</td>
<td>1.7058</td>
<td>0.5</td>
<td>1.3</td>
<td>46.7</td>
<td>0.7</td>
</tr>
<tr>
<td>61</td>
<td>Diapteron</td>
<td>Diapteron cyanosticum</td>
<td>17.24</td>
<td>1.7012</td>
<td>1.7016</td>
<td>0.4</td>
<td>1.6</td>
<td>42.4</td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>Diapteron</td>
<td>Diapteron cyanosticum + georgiae</td>
<td>17.07</td>
<td>1.7009</td>
<td>1.7016</td>
<td>0.7</td>
<td>2.2</td>
<td>42.4</td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>Epilampus</td>
<td>Epilampus chaperi</td>
<td>1.48</td>
<td>1.7010</td>
<td>1.7010</td>
<td></td>
<td></td>
<td></td>
<td>41.9</td>
</tr>
<tr>
<td>64</td>
<td>Pachypanchax</td>
<td>Pachypanchax playfairi</td>
<td>1.7046</td>
<td>1.7063</td>
<td>1.7063</td>
<td>1.7</td>
<td></td>
<td></td>
<td>47.2</td>
</tr>
<tr>
<td>65</td>
<td>Rivalus</td>
<td>Rivalus agilae</td>
<td>13.83</td>
<td>1.7007</td>
<td>1.7015</td>
<td>0.8</td>
<td>1.3</td>
<td>42.3</td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>Rivalus</td>
<td>Rivalus holmiae</td>
<td>18.12</td>
<td>1.6984</td>
<td>1.6987</td>
<td>0.3</td>
<td>1.8</td>
<td>39.5</td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>Cyprinodontidae</td>
<td>Fundulus heteroclitus</td>
<td>18.65</td>
<td>1.6993</td>
<td>1.6996</td>
<td>0.3</td>
<td>1.7</td>
<td>40.4</td>
<td>1.5</td>
</tr>
<tr>
<td>68</td>
<td>Cyprinodon</td>
<td>Cyprinodon macularius califor.</td>
<td>20.33</td>
<td>1.6989</td>
<td>1.6996</td>
<td>0.7</td>
<td>2.8</td>
<td>40.4</td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>Cyprinodon</td>
<td>Cyprinodon n. nevadensis</td>
<td>8.08</td>
<td>1.6983</td>
<td>1.6989</td>
<td>0.6</td>
<td>1.7</td>
<td>39.7</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>Cyprinodon</td>
<td>Cyprinodon salinus</td>
<td>10.21</td>
<td>1.6984</td>
<td>1.7000</td>
<td>1.6</td>
<td>4.4</td>
<td>40.8</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>Cyprinodon</td>
<td>Cyprinodon variegatus</td>
<td>7.54</td>
<td>1.6987</td>
<td>1.6998</td>
<td>1.1</td>
<td>3.1</td>
<td>40.6</td>
<td>1.6</td>
</tr>
<tr>
<td>72</td>
<td>Austrofundulus</td>
<td>Austrofundulus limnaei*</td>
<td>102.19</td>
<td>1.6971</td>
<td>1.6973</td>
<td>0.2</td>
<td>1.9</td>
<td>38.0</td>
<td></td>
</tr>
<tr>
<td>73</td>
<td>Jordanella</td>
<td>Jordanella floridai</td>
<td></td>
<td>1.6987</td>
<td>1.6987</td>
<td></td>
<td></td>
<td></td>
<td>39.5</td>
</tr>
<tr>
<td>74</td>
<td>Poeciliidae</td>
<td>Poecilia sphenops mel.</td>
<td>8.42</td>
<td>1.6987</td>
<td>1.6994</td>
<td>0.7</td>
<td>3.8</td>
<td>40.2</td>
<td>0.8</td>
</tr>
<tr>
<td>75</td>
<td>Xiphophorus</td>
<td>Xiphophorus maculatus</td>
<td>5.21</td>
<td>1.6985</td>
<td>1.6987</td>
<td>0.2</td>
<td>2.9</td>
<td>39.5</td>
<td>0.9</td>
</tr>
<tr>
<td>76</td>
<td>Syngnathiformes</td>
<td>Syngnathidae</td>
<td>25.66</td>
<td>1.7031</td>
<td>1.7038</td>
<td>0.7</td>
<td>3.0</td>
<td>44.7</td>
<td>0.7</td>
</tr>
<tr>
<td>77</td>
<td>Gobiosociformes</td>
<td>Gobiosocidae</td>
<td>40.00</td>
<td>1.6998</td>
<td>1.7002</td>
<td>0.3</td>
<td>4.0</td>
<td>41.0</td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>Dactylogiriformes</td>
<td>Dactylogiridae</td>
<td>9.84</td>
<td>1.7033</td>
<td>1.7040</td>
<td>0.7</td>
<td>3.9</td>
<td>44.9</td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>Scorpaeniformes</td>
<td>Scorpaenidae</td>
<td>33.25</td>
<td>1.6988</td>
<td>1.6990</td>
<td>0.2</td>
<td>4.2</td>
<td>39.8</td>
<td>1.4</td>
</tr>
<tr>
<td>80</td>
<td>Scorpaena</td>
<td>Scorpaena calarata</td>
<td>29.31</td>
<td>1.6995</td>
<td>1.6999</td>
<td>0.4</td>
<td>3.5</td>
<td>40.7</td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>Scorpaena</td>
<td>Scorpaena guttaea*</td>
<td>27.20</td>
<td>1.6974</td>
<td>1.6978</td>
<td>0.4</td>
<td>3.8</td>
<td>38.6</td>
<td></td>
</tr>
<tr>
<td>82</td>
<td>Sebastidae</td>
<td>Sebastidae atroirens*</td>
<td>22.53</td>
<td>1.6984</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>Triglidae</td>
<td>Prionotus carolinus</td>
<td>3.04</td>
<td>1.7011</td>
<td>1.7021</td>
<td>1.0</td>
<td>3.5</td>
<td>42.9</td>
<td>1.0</td>
</tr>
<tr>
<td>84</td>
<td>Cottidae</td>
<td>Hemirpipterus americanus</td>
<td>4.62</td>
<td>1.7010</td>
<td>1.7013</td>
<td>0.3</td>
<td>2.9</td>
<td>42.1</td>
<td>0.9</td>
</tr>
<tr>
<td>85</td>
<td>Leptocottus</td>
<td>Leptocottus armatus*</td>
<td>58.90</td>
<td>1.7012</td>
<td>1.7020</td>
<td>0.8</td>
<td>2.5</td>
<td>42.9</td>
<td></td>
</tr>
<tr>
<td>86</td>
<td>Cinocottus</td>
<td>Cinocottus analis*</td>
<td>63.00</td>
<td>1.7005</td>
<td>1.7004</td>
<td>-0.1</td>
<td>3.8</td>
<td>41.2</td>
<td>0.9</td>
</tr>
<tr>
<td>87</td>
<td>Hexagrammata</td>
<td>Ophiodyon elongatus*</td>
<td>21.00</td>
<td>1.6997</td>
<td>1.7006</td>
<td>0.9</td>
<td>3.9</td>
<td>41.5</td>
<td></td>
</tr>
<tr>
<td>Sample Order number</td>
<td>Family</td>
<td>Species</td>
<td>MW (kb)</td>
<td>$\rho_0$ (g/cm$^3$)</td>
<td>$\langle \rho \rangle$ (g/cm$^3$)</td>
<td>A (mg/cm$^3$)</td>
<td>H %GC</td>
<td>GC %</td>
<td>e (pg)</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------------</td>
<td>----------------------------------</td>
<td>--------</td>
<td>---------------------</td>
<td>------------------</td>
<td>--------------</td>
<td>------</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>88</td>
<td>Hexagrammidae</td>
<td>Pleurogrammus azonus*</td>
<td>11.56</td>
<td>1.6997</td>
<td>1.7004</td>
<td>0.7</td>
<td>3.4</td>
<td>41.2</td>
<td></td>
</tr>
<tr>
<td>89</td>
<td>Perciformes</td>
<td>Centropristis striatus</td>
<td>8.53</td>
<td>1.6981</td>
<td>1.6985</td>
<td>0.4</td>
<td>3.0</td>
<td>39.3</td>
<td>1.2</td>
</tr>
<tr>
<td>90</td>
<td>Serranidae</td>
<td>Epinephelus guttatus</td>
<td>10.09</td>
<td>1.7015</td>
<td>1.7013</td>
<td>-0.2</td>
<td>1.7</td>
<td>42.1</td>
<td>1.2</td>
</tr>
<tr>
<td>91</td>
<td>Serranidae</td>
<td>Epinephelus striatus</td>
<td>18.29</td>
<td>1.7004</td>
<td>1.7006</td>
<td>0.2</td>
<td>1.5</td>
<td>41.4</td>
<td>1.3</td>
</tr>
<tr>
<td>92</td>
<td>Paralabrax maculatofasciatus*</td>
<td></td>
<td>22.00</td>
<td>1.6979</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>93</td>
<td>Alphesines afer*</td>
<td>27.20</td>
<td>1.6988</td>
<td>1.6989</td>
<td>0.1</td>
<td>3.0</td>
<td>39.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>94</td>
<td>Paranthias colonus*</td>
<td>28.10</td>
<td>1.6997</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>Paralabrax clathratus*</td>
<td>23.60</td>
<td>1.6978</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>Acanthuridae</td>
<td>Acanthurus bahianus*</td>
<td>35.28</td>
<td>1.6990</td>
<td>1.6988</td>
<td>0.8</td>
<td>2.8</td>
<td>39.6</td>
<td></td>
</tr>
<tr>
<td>97</td>
<td>Acanthurus chirurgus*</td>
<td>43.00</td>
<td>1.6981</td>
<td>1.6985</td>
<td>0.3</td>
<td>2.5</td>
<td>39.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>98</td>
<td>Acanthurus coeruleus*</td>
<td>36.00</td>
<td>1.6979</td>
<td>1.6983</td>
<td>0.4</td>
<td>2.4</td>
<td>39.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>Acanthurus triostegus*</td>
<td>23.29</td>
<td>1.6989</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>Atherinidae</td>
<td>Leuresthes sardina*</td>
<td>46.60</td>
<td>1.6983</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>Cebidichthyidae</td>
<td>Cebidichthys violaceus*</td>
<td>55.00</td>
<td>1.6997</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>102</td>
<td>Chaetodontidae</td>
<td>Johnrandallia nigrostris*</td>
<td>24.30</td>
<td>1.6997</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>103</td>
<td>Echeneidae</td>
<td>Echeneis naucrates*</td>
<td>44.00</td>
<td>1.6993</td>
<td>1.69911</td>
<td>-0.2</td>
<td>2.8</td>
<td>40.1</td>
<td>0.7</td>
</tr>
<tr>
<td>104</td>
<td>Embiotocidae</td>
<td>Amphistichus argenteum*</td>
<td>27.06</td>
<td>1.6984</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>105</td>
<td>Brachyistius frenatus*</td>
<td>22.00</td>
<td>1.6988</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>106</td>
<td>Cymatogaster aggregata*</td>
<td>45.30</td>
<td>1.6990</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>107</td>
<td>Damalichthys vacca*</td>
<td>23.80</td>
<td>1.6985</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>108</td>
<td>Embiotoca jacksoni*</td>
<td>50.00</td>
<td>1.6990</td>
<td>1.6992</td>
<td>0.2</td>
<td>2.7</td>
<td>40.0</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>109</td>
<td>Embiotoca lateralis*</td>
<td>28.10</td>
<td>1.6989</td>
<td>1.6993</td>
<td>0.4</td>
<td>2.8</td>
<td>40.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>Hyperprosopon anale*</td>
<td>23.43</td>
<td>1.7002</td>
<td>1.7010</td>
<td>0.8</td>
<td>2.6</td>
<td>41.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>111</td>
<td>Hyperprosopon argenteum*</td>
<td>64.00</td>
<td>1.6992</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>112</td>
<td>Micrometrus minimus*</td>
<td>29.00</td>
<td>1.6991</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>113</td>
<td>Gobidae</td>
<td>Coryphopterus nicholsii*</td>
<td>36.60</td>
<td>1.6985</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>114</td>
<td>Gillichthys mirabilis*</td>
<td>20.80</td>
<td>1.6976</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>115</td>
<td>Gillichthys seta*</td>
<td>33.00</td>
<td>1.6984</td>
<td>1.6997</td>
<td>1.3</td>
<td>3.6</td>
<td>40.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>116</td>
<td>Zan Badge*</td>
<td>Oxybelus pictus*</td>
<td>30.21</td>
<td>1.7000</td>
<td>1.6997</td>
<td>-0.3</td>
<td>2.8</td>
<td>40.5</td>
<td></td>
</tr>
<tr>
<td>Sample Order number</td>
<td>Family</td>
<td>Species</td>
<td>MW (kb)</td>
<td>$\rho_0$ (g/cm³)</td>
<td>$&lt;\rho&gt;$ (g/cm³)</td>
<td>A (mg/cm³)</td>
<td>H (%GC)</td>
<td>GC (%)</td>
<td>e (pg)</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------</td>
<td>--------------------------------</td>
<td>---------</td>
<td>------------------</td>
<td>------------------</td>
<td>------------</td>
<td>---------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>117</td>
<td>Pomacanthidae</td>
<td>Holocanthus passer*</td>
<td>37.00</td>
<td>1.6999</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>118</td>
<td>Kyphosidae</td>
<td>Hermosilla azurea*</td>
<td>20.08</td>
<td>1.6997</td>
<td>1.7001</td>
<td>0.4</td>
<td>3.9</td>
<td>40.9</td>
<td></td>
</tr>
<tr>
<td>119</td>
<td>Labridae</td>
<td>Bodianus diploaena*</td>
<td>34.00</td>
<td>1.6997</td>
<td>1.6996</td>
<td>0.0</td>
<td>2.2</td>
<td>40.5</td>
<td></td>
</tr>
<tr>
<td>120</td>
<td></td>
<td>Bodianus rufus*</td>
<td>44.00</td>
<td>1.6990</td>
<td>1.6994</td>
<td>0.4</td>
<td>2.2</td>
<td>40.2</td>
<td></td>
</tr>
<tr>
<td>121</td>
<td></td>
<td>Clepticus parrae*</td>
<td>46.30</td>
<td>1.6990</td>
<td>1.6988</td>
<td>-0.1</td>
<td>2.7</td>
<td>39.6</td>
<td></td>
</tr>
<tr>
<td>122</td>
<td></td>
<td>Coris julis</td>
<td>22.70</td>
<td>1.7006</td>
<td>1.7015</td>
<td>0.9</td>
<td>2.9</td>
<td>42.3</td>
<td>1.2</td>
</tr>
<tr>
<td>123</td>
<td></td>
<td>Gomphosus varius*</td>
<td>36.00</td>
<td>1.6972</td>
<td>1.6977</td>
<td>0.5</td>
<td>2.5</td>
<td>38.4</td>
<td></td>
</tr>
<tr>
<td>124</td>
<td></td>
<td>Halichoeres garnot*</td>
<td>60.00</td>
<td>1.6985</td>
<td>1.6987</td>
<td>0.2</td>
<td>2.6</td>
<td>39.5</td>
<td></td>
</tr>
<tr>
<td>125</td>
<td></td>
<td>Halichoeres nicholsi*</td>
<td>30.00</td>
<td>1.6977</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>126</td>
<td></td>
<td>Halichoeres pictus*</td>
<td>40.50</td>
<td>1.6979</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>127</td>
<td></td>
<td>Halichoeres poeyi*</td>
<td>51.00</td>
<td>1.6986</td>
<td>1.6988</td>
<td>0.2</td>
<td>2.6</td>
<td>39.6</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td></td>
<td>Pseudodax moluccanus*</td>
<td>19.00</td>
<td>1.6995</td>
<td>1.7000</td>
<td>0.5</td>
<td>3.0</td>
<td>40.8</td>
<td></td>
</tr>
<tr>
<td>129</td>
<td></td>
<td>Semicyclus pulcher*</td>
<td>28.16</td>
<td>1.6995</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>130</td>
<td></td>
<td>Synodus cinereus</td>
<td>24.79</td>
<td>1.7002</td>
<td>1.7016</td>
<td>1.4</td>
<td>3.2</td>
<td>42.4</td>
<td>1.2</td>
</tr>
<tr>
<td>131</td>
<td></td>
<td>Synodus mediterraneus</td>
<td>17.94</td>
<td>1.7006</td>
<td>1.7015</td>
<td>0.9</td>
<td>3.1</td>
<td>42.3</td>
<td>0.6</td>
</tr>
<tr>
<td>132</td>
<td></td>
<td>Synodus ocellatus</td>
<td>20.14</td>
<td>1.7004</td>
<td>1.7012</td>
<td>0.8</td>
<td>3.2</td>
<td>42.0</td>
<td>1.1</td>
</tr>
<tr>
<td>133</td>
<td></td>
<td>Thalassoma amblycephalum*</td>
<td>31.80</td>
<td>1.6971</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>134</td>
<td></td>
<td>Thalassoma bifasciatum*</td>
<td>25.80</td>
<td>1.6982</td>
<td>1.6979</td>
<td>-0.2</td>
<td>3.5</td>
<td>38.7</td>
<td>1.0</td>
</tr>
<tr>
<td>135</td>
<td></td>
<td>Thalassoma grammaticum*</td>
<td>29.80</td>
<td>1.6974</td>
<td>1.6979</td>
<td>0.5</td>
<td>3.8</td>
<td>38.7</td>
<td></td>
</tr>
<tr>
<td>136</td>
<td></td>
<td>Thalassoma hardwicke*</td>
<td>62.00</td>
<td>1.6975</td>
<td>1.6976</td>
<td>0.1</td>
<td>2.5</td>
<td>38.3</td>
<td></td>
</tr>
<tr>
<td>137</td>
<td></td>
<td>Thalassoma hebraicum*</td>
<td>24.60</td>
<td>1.6970</td>
<td>1.6972</td>
<td>0.2</td>
<td>3.2</td>
<td>37.9</td>
<td></td>
</tr>
<tr>
<td>138</td>
<td></td>
<td>Thalassoma lucasanum*</td>
<td>25.00</td>
<td>1.6976</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>139</td>
<td></td>
<td>Thalassoma purpureum*</td>
<td>63.00</td>
<td>1.6976</td>
<td>1.6975</td>
<td>-0.2</td>
<td>2.3</td>
<td>38.2</td>
<td></td>
</tr>
<tr>
<td>140</td>
<td></td>
<td>Thalassoma trilobatum*</td>
<td>26.50</td>
<td>1.6971</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>141</td>
<td>Labrisomidae</td>
<td>Xyrichtys novacula</td>
<td>28.14</td>
<td>1.6998</td>
<td>1.7001</td>
<td>0.3</td>
<td>2.0</td>
<td>40.9</td>
<td></td>
</tr>
<tr>
<td>142</td>
<td>Pomatocentridae</td>
<td>Dialomus fuscus*</td>
<td>53.00</td>
<td>1.7001</td>
<td>1.7007</td>
<td>0.5</td>
<td>2.7</td>
<td>41.5</td>
<td></td>
</tr>
<tr>
<td>143</td>
<td></td>
<td>Chromis atrovirens*</td>
<td>27.00</td>
<td>1.6986</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>144</td>
<td></td>
<td>Chromis chromis</td>
<td>17.76</td>
<td>1.7005</td>
<td>1.6993</td>
<td>-1.2</td>
<td>4.4</td>
<td>40.1</td>
<td>1.2</td>
</tr>
<tr>
<td>145</td>
<td></td>
<td>Chromis cyanina*</td>
<td>17.57</td>
<td>1.6982</td>
<td>1.6987</td>
<td>0.5</td>
<td>5.2</td>
<td>39.5</td>
<td></td>
</tr>
<tr>
<td>Sample number</td>
<td>Family</td>
<td>Species</td>
<td>MW (kb)</td>
<td>$\rho_0$ (g/cm³)</td>
<td>$&lt;\rho&gt;$ (g/cm³)</td>
<td>A (mg/cm³)</td>
<td>H (%GC)</td>
<td>GC (%)</td>
<td>c (pg)</td>
</tr>
<tr>
<td>---------------</td>
<td>-------------------</td>
<td>--------------------------</td>
<td>---------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>------------</td>
<td>---------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td>146</td>
<td>Chromis multilineata*</td>
<td>38.50</td>
<td>1.6983</td>
<td>1.6984</td>
<td>0.1</td>
<td>3.1</td>
<td>39.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>147</td>
<td>Dascyllus aruanus*</td>
<td>32.40</td>
<td>1.6991</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>148</td>
<td>Dascyllus flavicanus*</td>
<td>29.90</td>
<td>1.6985</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>149</td>
<td>Microspathodon chrysops*</td>
<td>30.00</td>
<td>1.6979</td>
<td>1.6984</td>
<td>0.5</td>
<td>3.4</td>
<td>39.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>Microspathodon dorsalis*</td>
<td>35.00</td>
<td>1.6985</td>
<td>1.6989</td>
<td>0.4</td>
<td>4.1</td>
<td>39.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>151</td>
<td>Stegastes doriophanicos*</td>
<td>18.77</td>
<td>1.6990</td>
<td>1.6990</td>
<td>-0.1</td>
<td>3.8</td>
<td>39.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>152</td>
<td>Stegastes planifrons*</td>
<td>21.00</td>
<td>1.6991</td>
<td>1.6992</td>
<td>0.1</td>
<td>3.0</td>
<td>40.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>153</td>
<td>Paracarrihytidae</td>
<td>Paracarrihytiches forsteri</td>
<td>22.23</td>
<td>1.7006</td>
<td>1.7015</td>
<td>0.9</td>
<td>3.4</td>
<td>42.3</td>
<td></td>
</tr>
<tr>
<td>154</td>
<td>Scarididae</td>
<td>Calotomus catolimus*</td>
<td>56.00</td>
<td>1.6967</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>155</td>
<td>Hypposcarus harid*</td>
<td>39.00</td>
<td>1.6976</td>
<td>1.6980</td>
<td>0.4</td>
<td>2.6</td>
<td>38.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>156</td>
<td>Niclosina denticulata*</td>
<td>62.00</td>
<td>1.6965</td>
<td>1.6964</td>
<td>-0.1</td>
<td>3.0</td>
<td>37.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>157</td>
<td>Scarus coelestes*</td>
<td>78.00</td>
<td>1.6974</td>
<td>1.6976</td>
<td>0.2</td>
<td>2.5</td>
<td>38.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>158</td>
<td>Scarus ghobban*</td>
<td>77.00</td>
<td>1.6973</td>
<td>1.6982</td>
<td>0.9</td>
<td>2.9</td>
<td>39.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>159</td>
<td>Scarus gibbus*</td>
<td>12.79</td>
<td>1.6984</td>
<td>1.6988</td>
<td>0.4</td>
<td>1.6</td>
<td>39.6</td>
<td>1.9–2.1</td>
<td></td>
</tr>
<tr>
<td>160</td>
<td>Scarus psittacus*</td>
<td>50.00</td>
<td>1.6973</td>
<td>1.6977</td>
<td>0.4</td>
<td>2.7</td>
<td>38.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>161</td>
<td>Scarus schonegeli*</td>
<td>24.20</td>
<td>1.6970</td>
<td>1.6964</td>
<td>-0.6</td>
<td>3.7</td>
<td>37.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>162</td>
<td>Zanclusinae</td>
<td>Zanclus cornutus*</td>
<td>31.80</td>
<td>1.6975</td>
<td>1.6979</td>
<td>0.4</td>
<td>2.4</td>
<td>38.7</td>
<td></td>
</tr>
<tr>
<td>163</td>
<td>Lutjanidae</td>
<td>Lutjanus syngis</td>
<td>2.62</td>
<td>1.7010</td>
<td>1.7022</td>
<td>1.2</td>
<td>4.2</td>
<td>43.1</td>
<td>0.9–1.4</td>
</tr>
<tr>
<td>164</td>
<td>Leithrniidae</td>
<td>Leithrnius nebulosus</td>
<td>11.53</td>
<td>1.6999</td>
<td>1.7001</td>
<td>0.2</td>
<td>1.7</td>
<td>40.9</td>
<td></td>
</tr>
<tr>
<td>165</td>
<td>Cichlididae</td>
<td>Alocapra aliculae grahami</td>
<td>8.64</td>
<td>1.7007</td>
<td>1.7012</td>
<td>0.5</td>
<td>3.6</td>
<td>42.0</td>
<td></td>
</tr>
<tr>
<td>166</td>
<td>Cichlasoma neeki</td>
<td>1.7007</td>
<td>1.6994</td>
<td>-1.3</td>
<td></td>
<td></td>
<td></td>
<td>40.2</td>
<td>1.4</td>
</tr>
<tr>
<td>167</td>
<td>Oreochromis spilurus</td>
<td>6.15</td>
<td>1.7003</td>
<td>1.7006</td>
<td>0.3</td>
<td>2.2</td>
<td>41.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>168</td>
<td>Oreochromis aureus</td>
<td>9.11</td>
<td>1.7003</td>
<td>1.7008</td>
<td>0.5</td>
<td>2.3</td>
<td>41.6</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>169</td>
<td>Oreochromis mossambicus</td>
<td>9.71</td>
<td>1.7009</td>
<td>1.7010</td>
<td>0.1</td>
<td>3.0</td>
<td>41.8</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>170</td>
<td>Oreochromis niloticus</td>
<td>7.65</td>
<td>1.7002</td>
<td>1.7009</td>
<td>0.7</td>
<td>2.0</td>
<td>41.7</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>171</td>
<td>Symphysodon discus</td>
<td>24.36</td>
<td>1.7004</td>
<td>1.7005</td>
<td>0.1</td>
<td>2.5</td>
<td>41.3</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>172</td>
<td>Tilapia munikofei</td>
<td>25.88</td>
<td>1.7005</td>
<td>1.7009</td>
<td>0.4</td>
<td>2.5</td>
<td>40.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>173</td>
<td>Sphyraendidae</td>
<td>Sphyraena barracuda</td>
<td>8.64</td>
<td>1.7015</td>
<td>1.7017</td>
<td>0.2</td>
<td>1.8</td>
<td>42.6</td>
<td>0.8–1.2</td>
</tr>
<tr>
<td>174</td>
<td>Sphyraenidae</td>
<td>Sphyraena ensis</td>
<td>8.76</td>
<td>1.7024</td>
<td>1.7027</td>
<td>0.3</td>
<td>2.7</td>
<td>43.6</td>
<td></td>
</tr>
</tbody>
</table>
TABLE 4.2  
Properties of DNAs from fishes* (continued)

<table>
<thead>
<tr>
<th>Sample Order number</th>
<th>Family</th>
<th>Species</th>
<th>MW (kb)</th>
<th>(\rho_0) (g/cm(^3))</th>
<th>(\langle \rho \rangle) (g/cm(^3))</th>
<th>A (mg/cm(^3))</th>
<th>H (%GC)</th>
<th>GC (%)</th>
<th>(\epsilon) (pg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>175</td>
<td>Nototheniidae</td>
<td>Pagophila borchgrevinki</td>
<td>15.56</td>
<td>1.7002</td>
<td>1.7010</td>
<td>0.8</td>
<td>2.5</td>
<td>41.8</td>
<td></td>
</tr>
<tr>
<td>176</td>
<td>Trematomus bernacchii</td>
<td></td>
<td>4.25</td>
<td>1.7011</td>
<td>1.7020</td>
<td>0.9</td>
<td>3.1</td>
<td>42.9</td>
<td>1.9</td>
</tr>
<tr>
<td>177</td>
<td>Trematomus centronotus</td>
<td></td>
<td>1.82</td>
<td>1.7009</td>
<td>1.7017</td>
<td>0.8</td>
<td>4.1</td>
<td>42.5</td>
<td>2.0</td>
</tr>
<tr>
<td>178</td>
<td>Trematomus hansonii</td>
<td></td>
<td>13.38</td>
<td>1.6999</td>
<td>1.7002</td>
<td>0.3</td>
<td>2.4</td>
<td>41.1</td>
<td>1.8</td>
</tr>
<tr>
<td>179</td>
<td>Trematomus nicolai</td>
<td></td>
<td>4.71</td>
<td>1.7010</td>
<td>1.7017</td>
<td>0.7</td>
<td>3.7</td>
<td>42.5</td>
<td></td>
</tr>
<tr>
<td>180</td>
<td>Trematomus newnesi</td>
<td></td>
<td>1.53</td>
<td>1.7005</td>
<td>1.7018</td>
<td>1.3</td>
<td>4.5</td>
<td>42.6</td>
<td>2.1</td>
</tr>
<tr>
<td>181</td>
<td>Dissosticus mawsoni</td>
<td></td>
<td>9.84</td>
<td>1.7003</td>
<td>1.7007</td>
<td>0.4</td>
<td>2.5</td>
<td>41.5</td>
<td></td>
</tr>
<tr>
<td>182</td>
<td>Bathyrhinaeidae</td>
<td>Gymnomura acuticeps</td>
<td>17.24</td>
<td>1.7007</td>
<td>1.7012</td>
<td>0.5</td>
<td>3.5</td>
<td>42.1</td>
<td>1.9</td>
</tr>
<tr>
<td>183</td>
<td>Blemidae</td>
<td>Ophioleinius atlanticus</td>
<td>7.33</td>
<td>1.7007</td>
<td>1.7011</td>
<td>0.4</td>
<td>2.8</td>
<td>41.9</td>
<td>0.8–1.0</td>
</tr>
<tr>
<td>184</td>
<td>Calyonemidae</td>
<td>Synchirops splendidus</td>
<td>11.12</td>
<td>1.7073</td>
<td>1.7076</td>
<td>0.3</td>
<td>2.7</td>
<td>48.6</td>
<td></td>
</tr>
<tr>
<td>185</td>
<td>Pleuronectiformes</td>
<td>Pleuronectidae</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>186</td>
<td>Pleuronectidae</td>
<td>Siacum papillosum</td>
<td>2.62</td>
<td>1.7024</td>
<td>1.7025</td>
<td>0.1</td>
<td>4.7</td>
<td>43.4</td>
<td>0.7–1.0</td>
</tr>
<tr>
<td>187</td>
<td>Lepidobranchus aspera</td>
<td>Lepidobranchus aspera</td>
<td>6.62</td>
<td>1.6995</td>
<td>1.7002</td>
<td>0.7</td>
<td>3.5</td>
<td>41.0</td>
<td>0.6–1.0</td>
</tr>
<tr>
<td>188</td>
<td>Lepidobranchus ferrugmosa</td>
<td></td>
<td>10.34</td>
<td>1.7012</td>
<td>1.7021</td>
<td>0.9</td>
<td>4.7</td>
<td>42.9</td>
<td></td>
</tr>
<tr>
<td>189</td>
<td>Psettichthys melanostictus</td>
<td></td>
<td>80.00</td>
<td>1.7002</td>
<td>1.7006</td>
<td>0.3</td>
<td>2.7</td>
<td>41.4</td>
<td></td>
</tr>
<tr>
<td>190</td>
<td>Pleuronichthys californicus</td>
<td></td>
<td>49.00</td>
<td>1.6996</td>
<td>1.6992</td>
<td>–0.4</td>
<td>2.7</td>
<td>40.0</td>
<td></td>
</tr>
<tr>
<td>191</td>
<td>Tetraodontiformes</td>
<td>Balistidae</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>192</td>
<td>Balistidae</td>
<td>Melichthys vidua*</td>
<td>12.79</td>
<td>1.7014</td>
<td>1.7022</td>
<td>0.8</td>
<td>2.5</td>
<td>43.1</td>
<td></td>
</tr>
<tr>
<td>193</td>
<td>Balistidae</td>
<td>Balistes capriscus</td>
<td>17.07</td>
<td>1.7023</td>
<td>1.7024</td>
<td>0.1</td>
<td>2.4</td>
<td>43.3</td>
<td>0.7</td>
</tr>
<tr>
<td>194</td>
<td>Balistidae</td>
<td>Monacanthus tucker*</td>
<td>37.80</td>
<td>1.7042</td>
<td>1.7046</td>
<td>0.4</td>
<td>3.6</td>
<td>45.5</td>
<td></td>
</tr>
<tr>
<td>195</td>
<td>Balistidae</td>
<td>Rhinecanthus aculeatus</td>
<td>8.19</td>
<td>1.7047</td>
<td>1.7048</td>
<td>0.1</td>
<td>1.5</td>
<td>45.7</td>
<td></td>
</tr>
<tr>
<td>196</td>
<td>Balistidae</td>
<td>Stephanolpis hispidus</td>
<td>3.73</td>
<td>1.7037</td>
<td>1.7043</td>
<td>0.6</td>
<td>4.4</td>
<td>45.2</td>
<td>0.7</td>
</tr>
<tr>
<td>197</td>
<td>Ostracidae</td>
<td>Acanthodoradon quadricornis</td>
<td>51.10</td>
<td>1.7001</td>
<td>1.7008</td>
<td>0.7</td>
<td>2.7</td>
<td>41.6</td>
<td></td>
</tr>
<tr>
<td>198</td>
<td>Tetraodontidae</td>
<td>Lagoscepolis laevisatus</td>
<td>9.35</td>
<td>1.7022</td>
<td>1.7026</td>
<td>0.4</td>
<td>3.5</td>
<td>43.5</td>
<td>0.4–0.5</td>
</tr>
<tr>
<td>199</td>
<td>Tetraodontidae</td>
<td>Arothron diadematus</td>
<td>10.09</td>
<td>1.7025</td>
<td>1.7036</td>
<td>1.1</td>
<td>3.0</td>
<td>44.5</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>Tetraodontidae</td>
<td>Arothron meleagris</td>
<td>26.77</td>
<td>1.7034</td>
<td>1.7058</td>
<td>1.4</td>
<td>4.8</td>
<td>46.7</td>
<td></td>
</tr>
<tr>
<td>201</td>
<td>Tetraodontidae</td>
<td>Sphoeroides annulatus</td>
<td>18.47</td>
<td>1.7047</td>
<td>1.7053</td>
<td>0.6</td>
<td>3.6</td>
<td>46.2</td>
<td></td>
</tr>
</tbody>
</table>

*a Asterisks indicate values from Bucciarelli et al. (2002). All other values are from Bernardi and Bernardi (1990a). See this reference for additional information and for the sources of \(\epsilon\) values. The sources of samples are given in Bernardi and Bernardi (1990a) and in Bucciarelli et al. (2002). See also legend of Fig. 4.1.
heavy side and much more rarely on the light side. Finally, some other samples showed poorly resolved satellite on the heavy side.

Figure 4.2 displays the distribution of the modal buoyant densities of the fish DNAs studied. The modal buoyant density range covered was 1.696–1.708 g/cm³. In the case of Actinopterygii, the average $\rho_0$ was 1.7014 g/cm³ and the corresponding standard deviation, $\sigma$, was 2.2 mg/cm³; most values were between 1.699 and 1.704 g/cm³. In the case of Chondrichthyes, the average $\rho_0$ was 1.7035 g/cm³, and $\sigma$ was 1.1 mg/cm³; most values were in the 1.702–1.704 g/cm³ range. Because the distribution of $\rho_0$ values could be biased by the over-representation of species from the same genera or families of fishes, a histogram was constructed in which only one DNA sample per genus was taken into account when the other species within the genus had the same buoyant density. This histogram (Fig. 4.2A) was not, however, remarkably different from that including all species (see Bucciarelli et al., 2002).

The distribution of DNA CsCl band asymmetries for all fish species investigated are presented in the histogram of Fig. 4.2B. The majority of fishes showed asymmetry values around 0.5 mg/cm³ ($A = 0.65$ mg/cm³; $\sigma = 0.54$ mg/cm³ in the case of Actinopterygii).

Figure 4.1. CsCl band profiles of fish DNAs. See Table 3 of the original paper for the numbering of samples. Interestingly, even the very low molecular weight samples (see Table 4.2) yielded acceptable profiles, because of the presence of large-size molecules in DNA preparations and of the compositional homogeneity of fish DNAs. (From Bernardi and Bernardi 1990a).
Intermolecular compositional heterogeneities (H) basically reflect the spread of GC levels among different fragments in a given DNA preparation. (see Part 1, section 1.3) Fig. 4.2C displays the distribution of H values for different fish species. Values ranged from 0.5% GC for *Aphyosemion striatum* (sample 59) to 4.7% GC for *Arothron meleagris* (sample 199; the higher value of 4.9% for *Squatina dumerili*, sample 7, was due to a shoulder on the heavy side of the main band, indicative of a satellite). The average heterogeneity, H, of DNAs from *Osteichthyes* was equal to 2.6% GC (σ = 1.1% GC). Values higher than 3% GC were often due to poorly resolved satellites (see, for instance, sample 7). Artefactual reasons, like low molecular weight of DNA samples, may also account for some high values of H. A plot of CsCl band asymmetry versus intermolecular compositional heterogeneity (Fig. 4.2) shows a linear relationship with a significant slope (p < 0.001), stressing the contribution of asymmetry to heterogeneity (see Part 1, Section 1.5). In most cases, compositional heterogeneities were equal to, or lower than, those of comparable bacterial DNAs (see Figs. 3.10 and 3.11).

In conclusion, an almost two-fold increase in the number of fish DNAs investigated, reaching a total of 200 species, including species from a number of orders and families not previously explored, did not lead to any significant differences in the range and average values of modal buoyant densities, asymmetry and heterogeneity compared to previous

![Figure 4.2](image-url)
investigations (Bernardi and Bernardi, 1990a). This suggests that these values are unlikely to change with further increases in the fish samples explored.

1.2. Compositional properties: a Cs$_2$SO$_4$/BAMD analysis

Fractionation in Cs$_2$SO$_4$/BAMD was used to analyze in more detail some fish genomes. The analyses of DNA fractions, presented as histograms of relative DNA amounts versus modal buoyant densities in CsCl in Fig. 4.3, provided information on cryptic, or poorly detectable, satellites, which all showed up as GC-rich satellites. When not resolved from the main band, cryptic satellites were sometimes detected by restriction enzyme digestions. These analyses were done for different reasons.

(i) Some analyses were used to confirm the correlation proposed by Cuny et al. (1981; see also Part 7) between chromosomal banding patterns and compositional heterogeneity of DNA (Medrano et al., 1988). Anguilla anguilla and Epinephelus guttatus exhibit different degrees of asymmetry (1.3 and 0.2 mg/cm$^3$, respectively) and compositional heterogeneity (2.3% and 1.7% GC, respectively). These species were shown to be characterized by strong differences in Giemsa and Reverse banding of their metaphase chromosomes. Indeed, practically no banding pattern was found in E. guttatus, whereas a rather distinct pattern could be detected in A. anguilla. Another species, Labeo bicolor, exhibiting intermediate values of asymmetry (0.7 mg/cm$^3$) and heterogeneity (2.0% GC) showed a banding pattern of intermediate intensity and reciprocity. Interestingly, the other two fishes explored, Salmo gairdneri (or Onchorynchus mykiss) and Cyprinus carpio, also showed a remarkable

![Figure 4.3](image-url)

Figure 4.3. Histograms showing the relative amounts and the modal buoyant densities in CsCl of DNA fractions obtained by preparative Cs$_2$SO$_4$/BAMD density gradient centrifugation from some fish species. Black bars correspond to satellite components. The vertical broken line at 1.705 g/cm$^3$ is shown to provide a reference. Numbers in parentheses refer to Table 4.2. (From Bernardi and Bernardi 1990a).
compositional heterogeneity in their DNAs. Yet, the chromosomal banding patterns were very different. In *S. gairdneri* the compositional spectrum was similar to that of *A. anguilla* and, expectedly, the Giemsa banding pattern was quite distinct. In the case of *C. carpio*, which is characterized by one of the GC-poorest DNAs of all vertebrates, chromosomes are extremely compact. (ii) In the case of the genome of *Pagothenia* (formerly *Trematomus*) *borchgrevinki*, a typical Antarctic fish (sample 175), the CsCl bands for the six fractions obtained show (Hudson et al., 1980) that 85% of the DNA is in the range 1.700-1.702 g/cm³ (fractions 1-3) pointing to a very homogeneous genome. Among higher density fractions, the most abundant, fraction 5, showed, upon digestion with restriction enzyme HaeIII, three strong bands indicative of a satellite component. (iii) *Arothron diadematus* was investigated because the order to which it belongs (*Tetraodontiformes*) is characterized by a very small genome (c = 0.4-0.5 pg), in fact, the smallest among vertebrate genomes. In this case, renaturation kinetics showed that this DNA is the vertebrate DNA lowest in repeated sequences (Pizon et al., 1984, see Fig. 4.4). The very homogeneous *A. diadematus* DNA could be resolved into several components, characterized by buoyant densities of 1.700, 1.704, 1.708, 1.712 and 1.723 g/cm³, and representing 15%, 73%, 4%, 4% and 2.5%, respectively, of total DNA. The last component comprised a satellite DNA and ribosomal DNA. A family of interspersed repeats, possibly related to the *Alu* family of warm-blooded vertebrates, showed an extremely specific genomic distribution, being present only in the 1.708 g/cm³ component, which it matched in base composition (see Fig. 4.3). Later work on another tetraodontiform, *Takifugu rubripes*, showed that these small genomes are also characterized by short introns (Brenner et al., 1993).

![Figure 4.4. Reassociation kinetics of *A. diadematus* DNA. Results with *Escherichia coli* DNA (-----) are shown for comparison. The solid line through the experimental points (•) is the overall profile resulting from the analysis of kinetic classes (Form Pizon et al., 1984; see Soriano et al., 1981)](image)

### 1.3. Compositional properties: an analysis of long sequences

The very recent availability of long DNA sequences from three fishes, a cyprinid, *Brachydanio rerio* (*Danio* or zebrafish), and two *Tetraodontiformes*, *Takifugu rubripes* (*Fugu* or pufferfish) and *Tetraodon nigroviridis*, made it possible to push further the compositional...
analysis of fish genomes (Jabbari and Bernardi, 2003a) and to compare them with the compositional patterns of 20 kb segments from the human genome (in all cases, satellite and ribosomal sequences were absent from the sequences). The differences observed in Fig. 3.17 between the genomes from warm-blooded (human) and cold-blooded (Xenopus) are found again in the Danio/human comparison (Fig. 4.5). The case of the other two fish genomes is of special interest because the genomes from the most recent order Tetraodontiformes are exceptionally small (0.4 pg/haploid cell), very poor in repeated sequences (Pizon et al., 1984), characterized by short introns (Brenner et al., 1993) and GC-rich in most species (Bernardi and Bernardi, 1990a). In spite of these specific features, these genomes share with those of all other species an extremely narrow compositional distribution of large DNA segments compared to those of warm-blooded vertebrates (Fig. 3.18).

Two additional conclusions could be drawn from Fig. 4.5: (i) a positive asymmetry (i.e., a skew on the GC-rich side) seems, indeed, to be a property of main-band DNA and is not due to satellites, since the latter were excluded from the large DNA segments analyzed; this conclusion provides a final confirmation of what could be deduced 1) from the CsCl profiles; in fact, it was most unlikely that all the positive asymmetries of the CsCl profiles were systematically due to cryptic, GC-rich satellites; and 2) from the Cs2SO4/BAMD fractionation which showed (Fig. 4.3) that GC-rich satellites were not the only, or even the main, contributors to the asymmetric distributions of DNA molecules; (ii) a remarkable difference in average GC level was found between A. diadematus (Table 4.2) and Fugu, which showed values of 44.5% and 45.3%, on the one hand, and T. nigroviridis, which showed a value of 49%, on the other hand. This surprising difference for the genomes of fishes belonging to the same subfamily, Tetraodontinae (Tyler, 1980), will be discussed in Part 12.
1.4. Compositional properties of coding sequences and introns

This investigation could only be performed on Fugu and Danio, no annotation being available for Tetraodon. In view of the large, 9%, difference in average GC level of long DNA segments (Fig. 4.5), it was not surprising to find a 6% difference in the average GC level of coding sequences (Fig. 4.6A) and in that of GC3 (Fig. 4.6B).

The linear correlations between GC levels of introns and those of the corresponding coding sequences or GC3 showed differences which increased with increasing GC, intron values being, however, lower compared to coding sequence. In the comparison with GC3 values, differences were small for GC-poor genes, but they increased very much for GC-rich genes (not shown).

1.5. Compositional correlations

Fig. 4.6C shows the compositional correlation between GC3 and flanking sequences (20 kb in size) for Fugu. This was used to assess gene density (Fig. 5.7), by the approach developed for estimating gene density in the human genome (see Part 3). Interestingly, the GC levels of flanking sequences corresponding to the highest GC3 values are only about 47%, versus 55% in the case of human isochores (see Fig. 3.20).

![Figure 4.6](image-url)

Figure 4.6. Relative amounts of coding sequences from Danio and Fugu are plotted against their GC (A) and GC3 (B) levels. C. Plot of GC3 against 20 kb segments of Fugu genome. (From Jabbari and Bernardi, 2004a).
In the case of amphibians, the basic compositional properties (Thiery et al., 1976; Bernardi and Bernardi, 1990a) were similar to those just described for fishes (see Table 4.3 and Fig. 4.7). The modal buoyant densities of DNAs from anurans range from 1.6991 g/cm$^3$ for *Xenopus laevis* (sample 1 of Table 4.3), a strictly aquatic species from the ancient superorder *Pipoidei*, to 1.7013, 1.7028, and 1.7029 g/cm$^3$ for *Leptodactylus pentadactylus*, *Bufo paracnemis*, and *Rana sp.* (samples 2-4), respectively; these three species belong to the more recent, not strictly aquatic, superorder *Ranoidei*. The $\rho_0$ value of DNA from the urodele *Pleurodeles waltl* (sample 5) is equal to 1.7041 g/cm$^3$.

The CsCl profiles of DNAs from amphibians (Fig. 4.7) show no satellite bands; cryptic satellites may exist, however. For instance, in the case of *X. laevis*, the CsCl analysis of fractions from preparative centrifugation in Cs$_2$SO$_4$/Ag$^+$ showed (see Fig. 3.3) the presence of minor components banding at 1.705 and 1.706 g/cm$^3$ and of a minute amount of a 1.712 g/cm$^3$ component (Thiery et al., 1976).

On the other hand, the Cs$_2$SO$_4$/Ag$^+$ fractionation of *P. waltl* DNA revealed the presence of a component banding at 1.706 g/cm$^3$ and representing about 25% of DNA, whereas the majority of DNA bands at 1.704 g/cm$^3$ (Thiery et al., 1976). It is possible that the heavier component is a satellite, as no such component was found after HindII+III digestion (Macaya et al., 1976), as if it had been degraded to small fragments.

All amphibian DNAs are characterized by low asymmetries (ranging from 0.4 to 0.8 mg/
<table>
<thead>
<tr>
<th>Class</th>
<th>Amphibians</th>
<th>Order</th>
<th>Species</th>
<th>$S_{20,w}$</th>
<th>$\rho_0$</th>
<th>$&lt;\rho&gt;$</th>
<th>A</th>
<th>H</th>
<th>GC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anura</td>
<td>Family</td>
<td>1</td>
<td><em>Xenopus laevis</em></td>
<td>35.0</td>
<td>1.6991</td>
<td>1.6997</td>
<td>0.6</td>
<td>2.9</td>
<td>40.5</td>
</tr>
<tr>
<td></td>
<td>Pipidae</td>
<td>2</td>
<td><em>Leptodactylus pentadactylus</em></td>
<td>1.7013</td>
<td>1.7018</td>
<td>1.7036</td>
<td>0.8</td>
<td>42.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bufonidae</td>
<td>3</td>
<td><em>Bufo paracnemis</em></td>
<td>1.7028</td>
<td>1.7036</td>
<td>1.7033</td>
<td>0.4</td>
<td>2.9</td>
<td>44.5</td>
</tr>
<tr>
<td></td>
<td>Ranidae</td>
<td>4</td>
<td><em>Rana sp.</em></td>
<td>39.7</td>
<td>1.7029</td>
<td>1.7035</td>
<td>0.4</td>
<td>2.9</td>
<td>44.2</td>
</tr>
<tr>
<td>Caudata</td>
<td>Salamandridae</td>
<td>5</td>
<td><em>Pleurodeles waltli</em></td>
<td>27.1</td>
<td>1.7041</td>
<td>1.7047</td>
<td>0.6</td>
<td>3.0</td>
<td>45.6</td>
</tr>
</tbody>
</table>

*a* From Bernardi and Bernardi, 1990a

*b* Data from Thiery et al. (1976) H values from Bernardi and Bernardi (1990a)
cm$^3$), the skewness always being on the heavy side of the CsCl bands. Heterogeneities were equal to 2.9–3.0% GC in the three cases analyzed. Both asymmetries and heterogeneities of DNAs from amphibians are close to those exhibited by most fish DNAs. It is likely, however, that both sets of values for amphibians are overestimated because of the presence of GC-rich cryptic satellites, as in the case of fishes.
Extant reptiles comprise three main groups: (i) turtles and (ii) crocodiles, two homogeneous orders accounting for a relatively small number of species; and (iii) lepidosaurs, or squamates, consisting of lizards and snakes, which are morphologically more diversified and richer in species (see Olmo et al., 2002). Reptiles were studied in three steps (Thiery et al., 1976; Bernardi and Bernardi, 1990a; Hughes et al., 2002) involving a total of 2, 13 and 43 species, respectively. Because of differences in the classification of reptiles, in the experimental setup (Beckman Optima XL versus Spinco model E ultracentrifuges) and in the slightly different criteria used to define CsCl band borders (and therefore asymmetries and heterogeneities), the older results (Bernardi and Bernardi, 1990a) and the more recent ones (Hughes et al., 2002) will be presented separately. The advantage of this separate presentation is that each data set is perfectly coherent. Moreover, the older set was assessed exactly like the fish and amphibian sets investigated by Bernardi and Bernardi (1990a; see Chapters 1 and 2).

As far as the older results are concerned, DNAs from reptiles exhibited modal buoyant densities that were in the range 1.6992–1.7047 g/cm³ (Table 4.4A and Fig. 4.8A). Shoulders on the CsCl bands can be seen in at least three DNAs, those from *Tarentola mauretanica, Natrix maura* and *Bothrops neuwiedi* (samples 5, 9, 13 of Table 4.4A). Moreover, cryptic satellites might be responsible for the GC-rich tails exhibited by the CsCl band profiles of most DNAs.

The asymmetries of CsCl bands of DNAs from reptiles were often close to those shown by DNAs from amphibians. They were, however, larger in several cases (Table 4.4A). In five cases out of the six in which heterogeneity was determined, values were close to those found in DNAs from amphibians, with a higher value for *Testudo graeca* (sample 3; Table 4.4A). Again, the presence of cryptic heavy satellites may be responsible, at least in part, for the higher asymmetry and heterogeneity values of DNAs from reptiles. Such presence is suggested by the existence of components melting at high temperature in DNAs from some reptiles (Olmo, 1981). Asymmetry and heterogeneity of DNAs from reptiles are in the same range as those of most fish DNAs, with some exceptions.

A crocodile (*Crocodylus niloticus*) and a turtle (*T. graeca*), two GC-rich genomes (see Table 4.4A) were studied in more detail, namely by Cs₂SO₄/BAMD fractionation and CpG island analysis (Aissani and Bernardi, 1991a). Fractionation results showed the presence of sizable amounts of heavy fractions in both DNAs (see Fig. 5.9). Methylation data (Jabbari et al., 1997) and the analysis of a few coding sequences (Hughes et al., 1999) from these reptilian genomes will be presented in Parts 5 and 12, respectively.

The more recent results (Hughes et al., 2002; see Table 4.4B and Fig. 4.8B) are characterized by $p_0$ values comprised between 1.6974 and 1.7032 g/cm³ (covering a lower range than that previously reported), a by higher asymmetry, and heterogeneity values. An interesting point of the new data is that often several species from the same sub-family
<table>
<thead>
<tr>
<th>Order</th>
<th>Family</th>
<th>Species</th>
<th>$S_{20\text{-}w}$</th>
<th>$\rho_0$ g/cm$^3$</th>
<th>$&lt;\rho&gt;$ g/cm$^3$</th>
<th>A mg/cm$^3$</th>
<th>H %</th>
<th>GC %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crocodylia</td>
<td>Crocodylidae</td>
<td>1 Crocodylus niloticus</td>
<td>37.7</td>
<td>1.7046</td>
<td>1.7051</td>
<td>0.5</td>
<td>3.2</td>
<td>46.0</td>
</tr>
<tr>
<td>Chelonia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suborder</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cryptodira</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Squamata</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sauia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gekkonidae</td>
<td></td>
<td>4 Iguana iguana$^b$</td>
<td>20.5</td>
<td>1.7015</td>
<td>1.7022</td>
<td>0.7</td>
<td>2.7</td>
<td>43.1</td>
</tr>
<tr>
<td>Scincidae</td>
<td></td>
<td>5 Tarentola mauritana</td>
<td>23.9</td>
<td>1.7047</td>
<td>1.7057</td>
<td>1.0</td>
<td>2.7</td>
<td>46.6</td>
</tr>
<tr>
<td>Tefidae</td>
<td></td>
<td>6 Chalcides ocellatus</td>
<td>17.039</td>
<td>1.7050</td>
<td>1.7050</td>
<td>1.1</td>
<td></td>
<td>45.9</td>
</tr>
<tr>
<td>Varaunidae</td>
<td></td>
<td>7 Amelita amela</td>
<td>17.014</td>
<td>1.7018</td>
<td>1.7018</td>
<td>0.4</td>
<td></td>
<td>42.6</td>
</tr>
<tr>
<td>Ophida</td>
<td>Colubridae</td>
<td>8 Varanus griseus</td>
<td>17.024</td>
<td>1.7029</td>
<td>1.7029</td>
<td>0.5</td>
<td></td>
<td>43.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9 Natrix maura</td>
<td>23.1</td>
<td>1.7002</td>
<td>1.7018</td>
<td>1.6</td>
<td>3.4</td>
<td>42.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 Tomodon dorsatus</td>
<td>17.014</td>
<td>1.7015</td>
<td>1.7015</td>
<td>0.1</td>
<td></td>
<td>42.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11 Dromicus porphyryus s.</td>
<td>17.006</td>
<td>1.7007</td>
<td>1.7007</td>
<td>0.1</td>
<td></td>
<td>41.5</td>
</tr>
<tr>
<td></td>
<td>Viperidae</td>
<td>12 Cerastes cerastes</td>
<td>1.6995</td>
<td>1.7007</td>
<td>1.7007</td>
<td>1.2</td>
<td></td>
<td>41.5</td>
</tr>
<tr>
<td></td>
<td>Crotalidae</td>
<td>13 Bothrops neuviedi p.</td>
<td>1.6992</td>
<td>1.7029</td>
<td>1.7029</td>
<td></td>
<td></td>
<td>43.8</td>
</tr>
</tbody>
</table>

$^a$ From Bernardi and Bernardi, 1990a

$^b$ Data from Thiery et al. (1976). H values from Bernardi and Bernardi (1990a)

$^c$ Satellite band
were studied and showed different asymmetries of CsCl profiles. In this case, obviously the lowest values should be the most representative, since changes within a subfamily are most likely due to changes in satellite DNAs (or interspersed repeats).

Fig. 4.9 presents the asymmetry and the modal buoyant density values from Table 4.4B (when more than one species per genus or subfamily was available the lowest asymmetry values was used). As far as asymmetry is concerned, all values (except for one, close to the borderline) are below the lower boundary of asymmetry of mammals and birds, with many values below the upper boundary of fishes and amphibians. It should be noted that asymmetry values of Table 4.4B tend to be higher than those of Table 4.4A. In terms of modal buoyant density, 1.701 g/cm³ seems to be the borderline between snakes and turtles on the lower GC side and lizard and crocodiles on the upper side.
<table>
<thead>
<tr>
<th>Order Suborder</th>
<th>Superfamily Family</th>
<th>Subfamily Species</th>
<th>$\rho_0$ (g/cm$^3$)</th>
<th>$&lt;\rho&gt;$ (g/cm$^3$)</th>
<th>$&lt;\rho&gt;-\rho_0$ (mg/cm$^3$)</th>
<th>H (GC%)</th>
<th>MW (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Squamates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lacertilia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gekkota</td>
<td>Gekkonidae</td>
<td>Gekko gecko</td>
<td>1.7032</td>
<td>1.7046</td>
<td>0.14</td>
<td>5.2</td>
<td>16</td>
</tr>
<tr>
<td>Scincomorpha</td>
<td></td>
<td>Lacerta viridis</td>
<td>1.7019</td>
<td>1.7031</td>
<td>0.12</td>
<td>3.5</td>
<td>$&gt;60$</td>
</tr>
<tr>
<td>Lacertidae</td>
<td>Podarcis muralis</td>
<td>1.7019</td>
<td>1.7025</td>
<td>0.6</td>
<td></td>
<td>3.1</td>
<td>$&gt;60$</td>
</tr>
<tr>
<td>Diploglossa</td>
<td></td>
<td>Anguis fragilis</td>
<td>1.7019</td>
<td>1.7032</td>
<td>0.13</td>
<td>3.2</td>
<td>$&gt;60$</td>
</tr>
<tr>
<td>Anguidae</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uropeltidae</td>
<td>Uropeltinae</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serpentes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colubridae</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colubrinae</td>
<td>Boiga dendrophila</td>
<td>1.6997</td>
<td>1.7007</td>
<td>1.0</td>
<td></td>
<td>4.3</td>
<td>$&gt;60$</td>
</tr>
<tr>
<td>Elaphe obsoleta</td>
<td></td>
<td>1.6984</td>
<td>1.6993</td>
<td>0.9</td>
<td></td>
<td>3.7</td>
<td>$&gt;60$</td>
</tr>
<tr>
<td>Gonyosoma oxycephalum</td>
<td>1.6988</td>
<td>1.7005</td>
<td>1.7</td>
<td>5.1</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lampropeltis triangulum polyzona</td>
<td>1.6983</td>
<td>1.7011</td>
<td>2.8</td>
<td>6.8</td>
<td>36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhinocelius lecontei</td>
<td>1.6996</td>
<td>1.7015</td>
<td>1.9</td>
<td>6.0</td>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xenodontinae</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clelia rustica</td>
<td></td>
<td>1.6989</td>
<td>1.6995</td>
<td>0.6</td>
<td></td>
<td>2.8</td>
<td>$&gt;60$</td>
</tr>
<tr>
<td>Boidinae</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lampropeltis olivaceus</td>
<td>1.6995</td>
<td>1.7022</td>
<td>2.7</td>
<td>6.7</td>
<td>45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dipsadinae</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Placercis euryzona</td>
<td>1.7010</td>
<td>1.7014</td>
<td>0.4</td>
<td>3.5</td>
<td>54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sibon longifrenis</td>
<td>1.7005</td>
<td>1.7015</td>
<td>1</td>
<td>4.3</td>
<td>57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sibon nebulata</td>
<td>1.7000</td>
<td>1.7009</td>
<td>0.9</td>
<td>3.9</td>
<td>41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elapidae</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bungarinae</td>
<td>Aspindelaps lubricus</td>
<td>1.6984</td>
<td>1.7020</td>
<td>3.6</td>
<td>7.6</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Walterinnesia aegyptiaca</td>
<td>1.6983</td>
<td>1.6995</td>
<td>1.2</td>
<td>3.5</td>
<td>$&gt;60$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Notechidinae</td>
<td>Notechis scutatus</td>
<td>1.6980</td>
<td>1.7000</td>
<td>2</td>
<td></td>
<td>4.4</td>
<td>$&gt;60$</td>
</tr>
</tbody>
</table>
TABLE 4.4B
Properties of DNAs from reptiles* (continued)

<table>
<thead>
<tr>
<th>Order Suborder</th>
<th>Superfamily Family</th>
<th>Subfamily Species</th>
<th>$\rho_0$ (g/cm$^3$)</th>
<th>$\langle \rho \rangle$ (g/cm$^3$)</th>
<th>$\langle \rho \rangle - \rho_0$ (mg/cm$^3$)</th>
<th>H (GC%)</th>
<th>MW (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viperidae</td>
<td>Viperinae</td>
<td>Bitis gabonica gabonica</td>
<td>1.6982</td>
<td>1.7003</td>
<td>2.1</td>
<td>5.5</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vipera ammodobytes</td>
<td>1.6980</td>
<td>1.6991</td>
<td>1.1</td>
<td>3.5</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Crotalinae</td>
<td>Crotalus lepidus</td>
<td>1.6985</td>
<td>1.7006</td>
<td>2.1</td>
<td>5.5</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tropidolaemus wagleri</td>
<td>1.6995</td>
<td>1.7026</td>
<td>3.1</td>
<td>7.3</td>
<td>23</td>
</tr>
<tr>
<td>Henophidia</td>
<td>Boidae</td>
<td>Boa constrictor</td>
<td>1.6982</td>
<td>1.6997</td>
<td>1.5</td>
<td>3.5</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Pythonidae</td>
<td>Python molurus bivittatus</td>
<td>1.6974</td>
<td>1.6999</td>
<td>2.5</td>
<td>5.3</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Python curtus</td>
<td>1.6983</td>
<td>1.7005</td>
<td>2.2</td>
<td>6.2</td>
<td>18</td>
</tr>
<tr>
<td>Turtles</td>
<td>Cryptodires</td>
<td>Emydidae</td>
<td>Trachemys scripta elegans</td>
<td>1.7006</td>
<td>1.7028</td>
<td>2.2</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Trachemys scripta elegans</td>
<td>1.7007</td>
<td>1.7031</td>
<td>2.4</td>
<td>4.9</td>
<td>&gt;60</td>
</tr>
<tr>
<td></td>
<td>Pleurodires</td>
<td>Pelomedusidae</td>
<td>Pelusios subinger</td>
<td>1.7002</td>
<td>1.7026</td>
<td>2.4</td>
<td>5.6</td>
</tr>
<tr>
<td>Crocodiles</td>
<td>Alligatoridae</td>
<td>Alligator mississippiensis</td>
<td>1.7017</td>
<td>1.7037</td>
<td>1.9</td>
<td>4.5</td>
<td>&gt;60</td>
</tr>
<tr>
<td></td>
<td>Crocodylidae</td>
<td>Crocodylus acutus ♀</td>
<td>1.7022</td>
<td>1.7040</td>
<td>1.8</td>
<td>4.3</td>
<td>&gt;60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Crocodylus acutus ♂</td>
<td>1.7026</td>
<td>1.7044</td>
<td>1.8</td>
<td>4.5</td>
<td>&gt;60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Crocodylus niloticus</td>
<td>1.7021</td>
<td>1.7036</td>
<td>1.5</td>
<td>4.1</td>
<td>&gt;60</td>
</tr>
<tr>
<td>Birds</td>
<td>Phasianidae</td>
<td>Coturnix coturnix</td>
<td>1.6990</td>
<td>1.7018</td>
<td>2.8</td>
<td>6.4</td>
<td>&gt;60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gallus gallus</td>
<td>1.6996</td>
<td>1.7026</td>
<td>3.0</td>
<td>6.4</td>
<td>25</td>
</tr>
<tr>
<td>Amphibians</td>
<td>Ranidae</td>
<td>Rana esculenta</td>
<td>1.7024</td>
<td>1.7028</td>
<td>0.4</td>
<td>3.1</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Pipidae</td>
<td>Xenopus laevis</td>
<td>1.6973</td>
<td>1.6980</td>
<td>0.7</td>
<td>2.9</td>
<td>55</td>
</tr>
</tbody>
</table>

* The classification given for reptiles is that of the Reptile EMBL database (http://www.embl-heidelberg.de/~uetz/LivingReptiles.html). Sources of samples are given in the original paper. Asymmetry values in bold are the values used in Fig. 4.9, where only the lowest values from each subfamily or genus with more than one species were used. Some results with avian and amphibian DNAs are also presented. (From Hughes et al., 2002).
In conclusion, asymmetry values of CsCl profiles place several reptiles together with the other cold-blooded vertebrates. There is, however, a tendency in other cases to exhibit higher values compared to fishes and amphibians, which are, however, below the range of asymmetry values of warm-blooded vertebrates (see Fig. 4.9). The most remarkable feature of Fig. 4.9 is that within the sub-order with the largest number of subfamilies or genera, the sub-order of snakes, the range of asymmetry of CsCl profiles is much wider than those of entire classes of vertebrates, fishes/amphibians or mammals/birds. In other words, there is no similarity of asymmetry (and of compositional pattern) by common descent in the reptiles under consideration. This is a strong indication that compositional patterns developed independently in small taxons (subfamilies, genera; see also Part 12, Chapter 5).
Figure 4.9. Buoyant modal density is plotted against asymmetry for samples from Table 4.4B. Only the lowest asymmetry values from each subfamily or genus with more than one species were used. Snakes, closed circles; lizards, open circles; turtles, closed squares; crocodilians, open squares. The two horizontal broken lines represent the approximate upper levels of asymmetry for fishes and amphibians, and the approximate lower levels of asymmetry for mammals and birds, respectively. The red and the blue regions correspond to the asymmetry ranges of mammals/birds and fishes/amphibians, respectively. The yellow region corresponds to intermediate values. The vertical broken lines at 1.701 g/cm$^3$ separates snakes and turtles from lizards and crocodiles. (Modified from Hughes et al., 2002)
Table 4.5 shows a classification of birds and the species investigated. Eight species from eight different orders, including both paleognathous and neognathous birds, were explored, six of them by the Cs$_2$SO$_4$/BAMD approach followed by analytical centrifugation in CsCl. Avian DNAs showed a high degree of similarity in their isochore patterns. Major DNA components reached higher GC levels in avian DNAs (see Table 4.6 and Fig. 4.10) compared to mammalian DNAs; moreover, the former exhibited 10–12% of their DNAs having a buoyant density of 1.710 g/cm$^3$ or higher, whereas the latter only showed 6–10% of their DNAs in that buoyant density range (Kadi et al., 1993).

**TABLE 4.5**  
A classification of birds and the species investigated in the present work$^4$

<table>
<thead>
<tr>
<th>Superorder and Order</th>
<th>Family</th>
<th>Species investigated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paleognath</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Struthioniformes</td>
<td>Struthionidae</td>
<td>Struthio camelus</td>
</tr>
<tr>
<td>Rheiformes</td>
<td>Rheidae</td>
<td>Rhea americana</td>
</tr>
<tr>
<td>Neognath</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sphenisciformes</td>
<td>Spheniscidae</td>
<td>Spheniscus demersus</td>
</tr>
<tr>
<td>Anseriformes</td>
<td>Anatidae</td>
<td>Cuirina moschata</td>
</tr>
<tr>
<td>Galliformes</td>
<td>Plassanidae</td>
<td>Gallus gallus</td>
</tr>
<tr>
<td>Charadriiformes</td>
<td>Laridae</td>
<td>Larus argentatus</td>
</tr>
<tr>
<td>Columbiformes</td>
<td>Columbidae</td>
<td>Columba livia</td>
</tr>
<tr>
<td>Passeriformes</td>
<td>Ploceidae</td>
<td>Passer domesticus</td>
</tr>
</tbody>
</table>

$^4$ (From Kadi et al., 1993). The classification of birds is from Perrins and Middelton (1985), except for the introduction of superorders (Sibley and Monroe, 1990).

Interestingly, significant linear correlations were found between coding sequences (and their first, second and third codon positions), flanking regions (5' and 3'), and introns, as is the case in the human genome. These compositional correlations were not limited to GC levels but extended to individual bases. Finally, an analysis of coding sequences has confirmed a correlation among GC$_3$, GC$_2$, and GC$_1$ (Musto et al., 1999).
TABLE 4.6
Buoyant densities of avian DNAs

<table>
<thead>
<tr>
<th>DNA</th>
<th>$\rho_0$ g/cm$^3$</th>
<th>GC-rich DNA$^b$ %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ostrich</td>
<td>1.7001</td>
<td>9</td>
</tr>
<tr>
<td>Rhea</td>
<td>1.6999</td>
<td>9</td>
</tr>
<tr>
<td>Penguin</td>
<td>1.7006</td>
<td></td>
</tr>
<tr>
<td>Duck</td>
<td>1.6991</td>
<td>12.5</td>
</tr>
<tr>
<td>Chicken</td>
<td>1.6997$^b$</td>
<td></td>
</tr>
<tr>
<td>Seagull</td>
<td>1.6998$^c$</td>
<td>12.4</td>
</tr>
<tr>
<td>Pigeon</td>
<td>1.7007</td>
<td>13</td>
</tr>
<tr>
<td>Sparrow</td>
<td>1.7002$^d$</td>
<td></td>
</tr>
</tbody>
</table>

$^a$ From Kadi et al. (1993). See the original paper for additional information.

$^b$ % of DNA having a buoyant density of 1.710 g/cm$^3$ or higher (neglecting satellite and minor components).

---

Figure 4.10. Analytical CsCl density profiles of unfractionated DNA preparations. (From Kadi et al., 1993).
In contrast with avian DNA, mammalian DNAs showed differences in their compositional patterns. The first indication in this direction came from the early work of Thiery et al. (1976) which showed a lower degree of asymmetry for the DNAs from mouse and rat (as well as for those from hamster and garden dormouse) compared to those of other mammals (see Table 3.2). Investigations by Salinas et al. (1986) on the mouse genome and by Zerial et al. (1986a) on the human genome demonstrated important differences between the isochore patterns of these genomes. Indeed, the mouse genome lacked a GC-rich DNA component, H3, which was present in the human genome. Initially, it was difficult to decide whether this GC-richest component H3 was the result of a GC increase in the sequences forming a genome compartment (the isochore family H3), or of an invasion of GC-rich repetitive sequences, like Alu sequences (which are much more numerous in human than in mouse). The first explanation was unequivocally shown to be correct by comparing the compositional properties of orthologous genes from human and *Xenopus* (see Part 11) and by investigating the effect of removing (by masking) repeated sequences from the genome (see Part 6).

Table 4.7 presents a classification of mammals and the species investigated. Figs. 4.11 and 4.12 show the \( \rho_0 \) values and the CsCl profiles of the DNAs under consideration here and Table 4.8 lists the parameters obtained from such profiles. The work of Sabeur et al. (1993) on 20 species belonging to nine of the 17 eutherian orders demonstrated the existence of a “general distribution” in species belonging to eight mammalian orders. The human isochore pattern was typical of this general distribution. A “myomorph distribution” was found in *Myomorpha* (such as mouse and rat; see above), but not in the other rodent infraorders *Sciuromorpha* and *Histricomorpha*, which shared the “general distribution”. Two other distributions were found in a megachiropteran (but not in a microchiropteran, which, again, shared the general distribution) and in pangolin (a species from the only genus of the order *Pholidota*), respectively. The main difference between the general distribution and the other distributions was that the former contains sizable amounts, 6–10\%, of GC-rich isochores (detected as DNA molecules equal to, or higher than, 1.710 g/cm\(^3\) in buoyant density), which are scarce, or absent, in the other distributions. It should be stressed that while the “myomorph distribution” is now understood (see Part 11), the other special distributions have not been investigated further. We know, however, that they are different for the “myomorph distribution” (see Sabeur et al., 1993).
<table>
<thead>
<tr>
<th>Order, sub-, and infraorder</th>
<th>Family</th>
<th>Species investigated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monotremata</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marsupialia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insectivora</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Erinaceidae</td>
<td>1. Erinaceus europaeus Hedgehog</td>
</tr>
<tr>
<td></td>
<td>Soricidae</td>
<td>2. Crocidura russula Shrew</td>
</tr>
<tr>
<td></td>
<td>Talpidae</td>
<td>3. Talpa europaea Mole</td>
</tr>
<tr>
<td>Dermoptera</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chiroptera</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Megachiroptera</td>
<td>Pteropodidae</td>
<td>4. Pteropus sp. Fruit bat</td>
</tr>
<tr>
<td>Microchiroptera</td>
<td>Vespertilionidae</td>
<td>5. Myotis myotis Bat</td>
</tr>
<tr>
<td>Primates</td>
<td>Hominidae</td>
<td>6. Homo sapiens Man</td>
</tr>
<tr>
<td>Edentata</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pholidota</td>
<td>Manidae</td>
<td>7. Manis sp Pangolin</td>
</tr>
<tr>
<td>Lagomorpha</td>
<td>Leporidae</td>
<td>8. Oryctolagus cuniculus Rabbit</td>
</tr>
<tr>
<td>Rodentia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sciurognathi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sciurotomorpha</td>
<td>Scuridae</td>
<td>9. Sciurus vulgaris Squirrel</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10. Marmotta monax Woodchuck</td>
</tr>
<tr>
<td>Myomorpha</td>
<td>Muridae</td>
<td>11. Rattus norvegicus Rat</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12. Mus musculus Mouse</td>
</tr>
<tr>
<td></td>
<td>Cricetidae</td>
<td>13. Cricetus norvegicus Hamster</td>
</tr>
<tr>
<td></td>
<td>Spalacidae</td>
<td>14. Spalax sp Mole rat</td>
</tr>
<tr>
<td></td>
<td>Gliridae</td>
<td>15. Glis glis Dormouse</td>
</tr>
<tr>
<td>Histrionicognathi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caviomorpha</td>
<td>Caviidae</td>
<td>16. Cavia porcellus Guinea pig</td>
</tr>
<tr>
<td>Cetacea</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carnivora</td>
<td>Canidae</td>
<td>17. Canis familiaris Dog</td>
</tr>
<tr>
<td></td>
<td>Felidae</td>
<td>18. Felis domesticus Cat</td>
</tr>
<tr>
<td>Pinnipedia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tubulidentata</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proboscidea</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hyracoidea</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sirenia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perissodactyla</td>
<td>Equidae</td>
<td>19. Equus caballus Horse</td>
</tr>
<tr>
<td>Artiodactyla</td>
<td>Bovidae</td>
<td>20. Bos taurus Calf</td>
</tr>
</tbody>
</table>

*From Nowak and Paradiso (1983). Rodents are classified according to Colbert and Morales (1991).*
TABLE 4.8
Buoyant density properties of mammalian DNAs

<table>
<thead>
<tr>
<th>Order</th>
<th>Species</th>
<th>$\rho_0$ (g/cm$^3$)</th>
<th>$&lt;\rho&gt;$ (g/cm$^3$)</th>
<th>$&lt;\rho&gt;-\rho_0$ (mg/cm$^3$)</th>
<th>GC-rich DNA$^b$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insectivora</td>
<td>1. Hedgehog</td>
<td>1.7004</td>
<td>1.7035</td>
<td>3.1</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>2. Shrew</td>
<td>1.6976</td>
<td>1.7001</td>
<td>2.5</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>3. Mole</td>
<td>1.6987</td>
<td>1.7011</td>
<td>2.4</td>
<td>7</td>
</tr>
<tr>
<td>Chiroptera</td>
<td>4. Fruit Bat</td>
<td>1.6965</td>
<td>1.7002</td>
<td>3.7</td>
<td>&lt;1</td>
</tr>
<tr>
<td></td>
<td>5. Bat</td>
<td>1.7000</td>
<td>1.7024</td>
<td>2.4</td>
<td>10</td>
</tr>
<tr>
<td>Primates</td>
<td>6. Man</td>
<td>1.6984</td>
<td>1.7006</td>
<td>2.2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>7. Pangolin</td>
<td>1.6997</td>
<td>1.7026</td>
<td>2.9</td>
<td>&lt;1</td>
</tr>
<tr>
<td>Pholidota</td>
<td>8. Rabbit</td>
<td>1.6996</td>
<td>1.7039</td>
<td>4.1</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>9. Squirrel</td>
<td>1.6982</td>
<td>1.7006</td>
<td>2.4</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>10. Woodchuck</td>
<td>1.6983</td>
<td>1.7026</td>
<td>4.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11. Rat</td>
<td>1.7008</td>
<td>1.7014</td>
<td>1.6</td>
<td>&lt;1</td>
</tr>
<tr>
<td></td>
<td>12. Mouse</td>
<td>1.7006</td>
<td>1.7009</td>
<td>0.3$^c$</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>13. Hamster</td>
<td>1.7000</td>
<td>1.7021</td>
<td>2.1</td>
<td>&lt;4</td>
</tr>
<tr>
<td></td>
<td>14. Mole Rat</td>
<td>1.7000</td>
<td>1.7019</td>
<td>1.9</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>15. Dormouse</td>
<td>1.6991</td>
<td>1.7004</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16. Guinea Pig</td>
<td>1.6989</td>
<td>1.7018</td>
<td>2.9</td>
<td>6</td>
</tr>
<tr>
<td>Carnivora</td>
<td>17. Cat</td>
<td>1.6986</td>
<td>1.7019</td>
<td>3.3</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>18. Dog</td>
<td>1.6990</td>
<td>1.7017</td>
<td>2.7</td>
<td>6</td>
</tr>
<tr>
<td>Perissodactyla</td>
<td>19. Horse</td>
<td>1.6990</td>
<td>1.7030</td>
<td>4.0</td>
<td>8</td>
</tr>
<tr>
<td>Artiodactyla</td>
<td>20. Calf</td>
<td>1.7000</td>
<td>1.7039</td>
<td>3.9</td>
<td>7</td>
</tr>
</tbody>
</table>

$^a$ From Sabeur et al (1993). See the original paper for additional information.

$^b$ This column lists rounded-out figures from Table 3 of Sabeur et al. (1993). Values equal to, or higher than 6% (typical of the “general distribution”) are shown in bold type.

$^c$ The light satellite DNA was not taken into account in calculating $<\rho>$.

---

Figure 4.11. Modal buoyant density ($\rho_0$) values for mammalian DNAs. For the numbering of samples see Table 4.8. (From Sabeur et al, 1993)
Buoyant density profiles of high-molecular-weight DNAs centrifuged in CsCl gradients, \textit{i.e.}, compositional distribution of 50- to 100-kb genomic fragments, have revealed a clear difference between the murids and most other mammals, including other rodents. Sequence analyses have revealed other, related, compositional differences between murids and nonmurids. The study of CsCl profiles of 17 rodent species representing 13 families showed that the modal buoyant densities obtained for rodents span the full range of values observed in other eutherians (Figs. 4.11 and 4.13; Table 4.9). Scatterplots of these and related CsCl profile parameters show groups of rodent families that agree largely with established rodent taxonomy, in particular with the monophyly of the \textit{Geomyoidea} superfamily and the position of the \textit{Dipodidae} family within the \textit{Myomorpha}. In contrast, while confirming and extending previously reported differences between the profiles of \textit{Myomorpha} and those of other rodents, the CsCl data question a traditional hypothesis positing \textit{Gliridae} within \textit{Myomorpha}, as does the recently sequenced mitochondrial genome of dormouse (Reyes et al., 1998).
Figure 4.13. CsCl profiles of 18 rodent species (solid line) and their best fits by broadened exponential distributions (thin dashed lines). The abscissa shows modal buoyant density in g/cm$^3$; the ordinate shows relative amounts of DNA. All profiles are normalized to unit area and shown in three columns in order of increasing modal buoyant density. No fits are shown for the two Geomyoidea profiles (see the original paper). The dashed vertical line, at 1.7015 g/cm$^3$, indicates the mean buoyant density of the 18 species (From Douady et al., 2000).
TABLE 4.9
Buoyant densities of rodent DNAs*

<table>
<thead>
<tr>
<th>Family</th>
<th>Subfamily</th>
<th>Species</th>
<th>$\rho_0$ (g/cm$^3$)</th>
<th>$&lt;\rho&gt;$ (g/cm$^3$)</th>
<th>$&lt;\rho&gt;-\rho_0$ (mg/cm$^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dipodidae</td>
<td>Arvicolinae</td>
<td><em>Allactaga severtzi</em></td>
<td>1.7003</td>
<td>1.7016</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Clethrionomys glareolus</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Clethrionomys nagerilglareolus</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Muridae</td>
<td><em>Microtus agrestis</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Microtus/Chionomys nivalis</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Arvicolia amphibius</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Murinae</td>
<td><em>Mus musculus</em></td>
<td>1.7007</td>
<td>1.7012</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Mus spretus</em></td>
<td>1.6996</td>
<td>1.7001</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Rattus norvegicus</em></td>
<td>1.7008</td>
<td>1.7024</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Rattus rattus</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Acomys minous</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Apodemus mystacinus</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Apodemus sylvaticus</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Apodemus flavicolli</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Apodemus agrarius</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cricetinae</td>
<td><em>Cricetulus sp.</em></td>
<td>1.7000</td>
<td>1.7021</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Spalax sp.</em></td>
<td>1.7000</td>
<td>1.7019</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>Sigmodontinae</td>
<td><em>Ocymys paricola</em></td>
<td>1.6992</td>
<td>1.6998</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Peromyscus maniculatus</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Heteromyidae</td>
<td><em>Heteromys gaimard</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Dipodomys ordii</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Dipodomys microps</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Dipodomys agilis</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Dipodomys panamintinus</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Dipodomys stephani</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Dipodomys ingens</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Dipodomys heermannii</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Dipodomys merrimii</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Dipodomys elator</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family</td>
<td>Subfamily</td>
<td>Species</td>
<td>$\rho_0$ (g/cm$^3$)</td>
<td>$&lt;\rho&gt;$ (g/cm$^3$)</td>
<td>$&lt;\rho&gt;-\rho_0$ (mg/cm$^3$)</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------------</td>
<td>--------------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>Geomyidae</td>
<td></td>
<td>Dipodomys nitratoides</td>
<td>&quot;</td>
<td>[1.698]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dipodomys spectabilis</td>
<td>&quot;</td>
<td>[1.700]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dipodomys deserti</td>
<td>&quot;</td>
<td>[1.701]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thomomys bottae</td>
<td>Pocket gopher</td>
<td>1.7006</td>
<td>1.7042</td>
<td>3.6</td>
</tr>
<tr>
<td>Gliidae (Myoxidae)</td>
<td>Gliinae (Myoxinae)</td>
<td>Glis glis (Myoximus glis)</td>
<td>Fat dormouse</td>
<td>1.6991</td>
<td>1.7004</td>
</tr>
<tr>
<td></td>
<td>Leithiinae</td>
<td>Eliomys quercinus</td>
<td>Garden dormouse</td>
<td>1.6981</td>
<td>1.7005</td>
</tr>
<tr>
<td>Sciuridae</td>
<td>Scirinae</td>
<td>Sciurus vulgaris</td>
<td>Red squirrel</td>
<td>1.6982</td>
<td>1.7006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Marmota monax</td>
<td>Marmot / woodchuck</td>
<td>1.6983</td>
<td>1.7023</td>
</tr>
<tr>
<td>Aplodontidae</td>
<td></td>
<td>Aplodontia rufo</td>
<td>Mountain beaver</td>
<td>1.6978</td>
<td>1.7016</td>
</tr>
<tr>
<td>Anomalaridae</td>
<td>Anomalarinae</td>
<td>Anomalarus sp</td>
<td>Scaly-tailed flying squirrels</td>
<td>1.6974</td>
<td>1.7013</td>
</tr>
<tr>
<td>Ctenodactylida</td>
<td></td>
<td>Ctenodactylus vali*</td>
<td>Thomas' gundi</td>
<td>1.6993</td>
<td>1.7016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Massoutiera nzabi</td>
<td>Mzab gundi</td>
<td>1.6989</td>
<td>1.7031</td>
</tr>
<tr>
<td>Echimyidae</td>
<td>Eumysopinae</td>
<td>Proechymys guayannensis</td>
<td>Terrestrial spiny rat</td>
<td>1.6990</td>
<td>1.7017</td>
</tr>
<tr>
<td>Caviidae</td>
<td>Cavinae</td>
<td>Cavia porcellus</td>
<td>Guine-pig</td>
<td>1.6989</td>
<td>1.7018</td>
</tr>
<tr>
<td>Dasypodidae</td>
<td></td>
<td>Dasypodinae</td>
<td>(Red-rumped) agouti</td>
<td>1.6976</td>
<td>1.7023</td>
</tr>
<tr>
<td>(Agoutidae)</td>
<td></td>
<td>Mysoprotactus acouchy</td>
<td>(Red) acouchy</td>
<td>1.6976</td>
<td>1.7017</td>
</tr>
<tr>
<td>Thrionomyidae</td>
<td></td>
<td>Thrionomys swinderianus</td>
<td>Cane rat</td>
<td>1.6988</td>
<td>1.7018</td>
</tr>
<tr>
<td>Erethizontidae</td>
<td></td>
<td>Coendou prehensilis</td>
<td>Pre-hensile tailed porcupine</td>
<td>1.6984</td>
<td>1.7011</td>
</tr>
<tr>
<td>Chinchillidae</td>
<td></td>
<td>Sphiggurus insidiosus</td>
<td>Hairy dwarf porcupine</td>
<td>1.6984</td>
<td>1.7021</td>
</tr>
</tbody>
</table>

* From Douady et al. (2000). Taxonomy of families and subfamilies follows Wilson and Reeder (1993, pp. 7–8). Sources are denoted by a, Douady et al. (2000); b, Sabeur et al. (1993); c, Mazrimas and Hatch (1972); d, Walker (1968) and Hennig and Walker (1970). Modal buoyant densities from sources c and d are approximate only and are shown in brackets, since they were estimated from published pictures of smoothed CsCl profiles; means and other parameters could not be calculated from those profiles because of nonlinear responses (unreliable vertical scales). Two species with lower molecular weights are denoted by asterisks.
Figure 4.14. Modal buoyant density and GC values of rodent and other eutherian DNAs. Closed and open symbols refer to the data of Douady et al. (2000) and those of Sabeur et al. (1993), respectively. Triangles, squares, circles and diamonds denote caviomorphs, muroids, other rodents, and other mammals, respectively. Each row shows the values grouped at the family or subfamily level. The systematic arrangement follows McKenna and Bell (1997). The grouping of the Ctenodactylidae (Sciruravida suborder) with the Hystricognatha suborder (leftmost bracket) is supported by results from nuclear DNA sequence comparisons (von Willebrand factor gene; Huchon et al., 2000). For clarity, the two identical values observed in each of the Erethizontidae (1.6984 g/cm³) and Dasyproctidae (1.6976 g/cm³) have been artificially separated. With one exception (from Leithiinae), all the Myomorpha studied have modal buoyant density, $p_0$, equal to 1.699 g/cm³ (dashed line); only one other rodent species (from Ctenodactylidae) has $p_0$ in this range. Other Eutheria are, from left to right: Pteropus sp. (megabat), Crocidura russula (shrew), Homo sapiens (man), Felis domesticus (cat), Talpa europaea (mole), Canis familiaris (dog), Manis sp. (Asian pangolin), Oryctolagus cuniculus (rabbit), Bos taurus (cow), and Erinaceus europaeus (hedgehog). (1) Not named or not appropriate; (2) also called Myoxidae; (3) also called Dasyproctidae.
Part 5

Sequence distribution in the vertebrate genomes
CHAPTER 1

Gene distribution in the vertebrate genome

1.1. The distribution of genes in the human genome: the two gene spaces

The first indication that coding sequences are unevenly distributed in the human genome came from the observation that the majority of the sequences first tested were localized in the GC-richest isochore of the H3 family (Bernardi et al., 1985b), which only represents 4-5% of the genome. A quantitative assessment could be made later, on 1,610 (Mouchiroud et al., 1991) and 4,270 (Zoubak et al., 1996) genes, using the following approach. The correlation of GC3 levels of coding sequences with the GC levels of the isochore in which genes were located (see Fig. 3.20A) allows positioning of the GC3 histogram of human genes relative to the distribution of DNA molecules, namely the CsCl profile. Once this positioning is done, the ratio of each histogram bar by the corresponding point on the CsCl profile provides the relative concentration of genes (Mouchiroud et al., 1991; Zoubak et al., 1996). Fig. 5.1A shows that gene concentration increases from a very low average level

![Figure 5.1 A](image-url)  
Figure 5.1 A. Profile of gene concentration (red dots) in the human genome, as obtained by dividing the relative numbers of genes in each 1.5% GC3 interval of the histogram of gene distribution (yellow bars) by the corresponding relative amounts of DNA deduced from the CsCl profile (blue line). Values beyond the last one are unreliable because the corresponding DNA amounts are very low. The positioning of the GC3 histogram relative to the CsCl profile is based on the correlation of GC3 vs. GC of the isochore embedding the corresponding genes shown in Fig. 3.20A. (Modified from Bernardi, 2000a.) B. Density of genes in isochore family. Relative numbers of sequences over relative amounts of isochore family are presented (from Zoubak et al., 1996).
in L1 isochores up to a 20-fold higher level in H3 isochores. Fig. 5.1A presents the data of Fig. 5.1B in the form of a gene density histogram.

Since it had been tacitly assumed until then that genes were uniformly distributed in eukaryotic genomes, the strikingly non-uniform gene distribution in the human genome (and, for that matter, in the genomes of all vertebrates; see below) came as a big surprise. Moreover, a closer look at the gene concentration line of Fig. 5.1A reveals a clear break in the slopes of the GC-poor and the GC-rich isochores at 60% GC of coding sequences and at 46% GC of isochores. The break defines two gene spaces in the human genome: a GC-rich, gene-rich space corresponding to the H2 and H3 isochores and representing about 12% of the genome, which was called the genome core (Bernardi, 1993a), and a GC-poor, gene-poor space corresponding to the L1, L2 and H1 isochores and representing about 88% of the genome, which was called the genome desert or the empty quarter (from the classical name of the Arabian desert; Bernardi, 2000a). About half of the human genes are located in the small genome core, the other half being located in the large genome desert. The properties of these two gene spaces are very distinct (see the following section).

The plot of gene density as a function of GC in the human genome, as determined for 9,315 genes by Lander et al. (2001) is essentially identical (see Fig. 5.2) to that published by us for 1,610 and 4,270 genes ten and five years before, respectively (Mouchiroud et al., 1991; Zoubak et al., 1996). Most surprisingly, Lander et al. (2001) did not mention our previous results, in spite of the fact that they were reported again in a review paper (Bernardi, 2000a) quoted in their reference list. As a consequence, some authors, apparently not familiar with the literature, are now quoting the paper by Lander et al. (2001) as the original report about the non-uniform gene distribution in the human genome. (Fortunately, this mistake was not made again by the Mouse Genome Sequencing Consortium, 2002). Interestingly, we observed that the gene density values of Lander et al. (2001) fall on

Figure 5.2. Gene density as a function of GC content. Values are less accurate at higher GC levels because the denominator is small. For 9,315 known genes mapped to the draft genome sequence, the local GC content was calculated in a window covering either the whole alignment or 20,000 bp centered around the midpoint of the alignment, whichever was larger. Ns (undefined nucleotides) in the sequence were not counted. GC content for the genome was calculated for adjacent nonoverlapping 20,000 bp windows across the sequence. Both the gene and the genome distribution were normalized to sum one. The two slopes added here to the original figure concern the genes from GC-poor and GC-rich isochores, respectively (Modified from Fig. 36b of Lander et al., 2001; from Bernardi, 2001).
two straight lines crossing each other at about 46% GC (see Fig. 5.2). Along the same line, an independent assessment of gene density in chromosomes 21 and 22 (Saccone et al., 2001) also showed two slopes crossing each other at about 45% GC (see Part 7). The explanation given by Lander et al. (2001) for the results of Figs. 5.1 and 5.2 is that “the correlation between gene density and GC appears to be due primarily to intron size, which drops markedly with increasing GC content” (see Fig. 5.3). This is an astounding explanation, since introns only represent 4-6% of the genome and cannot possibly account for the correlation.

1.2. Properties of the two gene spaces

The “genome core” is endowed with several specific features, which are quite distinct from those of the “genome desert”. Table 5.1 summarizes the properties of the two gene spaces and provides indications as to where they are presented and/or discussed in this book. Here we will only comment on some points concerning gene expression.

As reviewed by Pesole et al. (2001), regulation of gene expression is achieved through a series of complex mechanisms which can be basically divided into two distinct steps. The first one involves the control of transcription mediated by cis-acting DNA elements such as promoters, enhancers, locus control regions and silencers to produce a mature mRNA. This mechanism has been well characterized for many genes. The second step covers the post-transcriptional control of mRNA nucleo-cytoplasmic transport, translation efficiency, subcellular localization and stability. This step has been less comprehensively characterized, although it is known to be mediated by cis-acting RNA elements generally located in 5’ and 3’ mRNA untranslated regions (5’UTRs and 3’UTRs; Sonenberg, 1994; McCarthy and Kollmus, 1995; Pesole et al., 1997; 2000; Bashirullah et al., 1998; van der

<table>
<thead>
<tr>
<th>Gene space:</th>
<th>Genome core</th>
<th>Genome desert</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome, %</td>
<td>~ 12%</td>
<td>~ 88%</td>
<td>Fig. 3.10</td>
</tr>
<tr>
<td>Genes, %</td>
<td>~ 50%</td>
<td>~ 50%</td>
<td>Fig. 5.1</td>
</tr>
<tr>
<td>Gene density</td>
<td>High</td>
<td>Low</td>
<td>Figs. 5.1, 5.2, 5.9</td>
</tr>
<tr>
<td>CpG</td>
<td>High</td>
<td>Low</td>
<td>Fig. 5.12</td>
</tr>
<tr>
<td>Methylation</td>
<td>High</td>
<td>Low</td>
<td>Fig. 5.15</td>
</tr>
<tr>
<td>CpG islands</td>
<td>Abundant</td>
<td>Scarce</td>
<td>Fig. 5.9</td>
</tr>
<tr>
<td>Introns</td>
<td>Short</td>
<td>Long</td>
<td>Fig. 5.3, *</td>
</tr>
<tr>
<td>Untranslated sequences</td>
<td>Short</td>
<td>Long</td>
<td>*</td>
</tr>
<tr>
<td>Consensus start sequence</td>
<td>RCC atg R</td>
<td>AAR atg R</td>
<td>*</td>
</tr>
<tr>
<td>Replication</td>
<td>Early</td>
<td>Late</td>
<td>Part 7</td>
</tr>
<tr>
<td>Recombination</td>
<td>High</td>
<td>Low</td>
<td>Part 7, *</td>
</tr>
<tr>
<td>Transcription</td>
<td>Strong</td>
<td>Weak</td>
<td>*</td>
</tr>
<tr>
<td>Chromatin</td>
<td>Open</td>
<td>Closed</td>
<td>Part 7, *</td>
</tr>
<tr>
<td>Chromosominal location</td>
<td>Distal</td>
<td>Proximal</td>
<td>Part 7</td>
</tr>
</tbody>
</table>

* See this Section
Velden and Thomas, 1999). Unlike DNA-mediated regulatory signals whose activity is essentially associated with their primary structure, the biological activity of regulatory patterns acting at the RNA level relies on a combination of primary and secondary structure elements assembled in a consensus structure generally recognized by specific RNA-binding proteins.

(i) Introns. Small numbers and sizes of introns are generally viewed as advantageous features for genes that are transcribed in an extensive way. The short and relatively rare introns of the highly expressed genes which are preferentially located in GC-rich isochores (see Part 7) fit with this view. In the case of GC-poor genes, the abundance and the large size of introns would be favourable for alternative splicing, an important mechanism of expression regulation in tissue-specific genes (Bell et al., 1998).

Although apparently not noticed by Lander et al. (2001), their results on intron size (see Fig. 5.3) confirm the existence of two classes of genes characterized by distinctly different intron sizes (Carels and Bernardi, 2000) in that they show a sharp transition with a midpoint at about 45% GC, again in agreement with the boundary between the genome core and the empty quarter.

(ii) Transcription. Ten years ago it was proposed that "the H3 isochore family presumably has the highest level of transcription because of its very high concentration of genes, especially housekeeping genes" (Bernardi, 1993a). Subsequent analyses on the correlation between gene distribution and gene expression levels (Gonçalves et al., 2000) claimed that the majority of the widely expressed (housekeeping) genes were localized mainly in GC-poor isochores, whereas tissue-specific genes were localized in the GC-rich ones. The authors drew this conclusions by analyzing the base composition and the distribution of genes with or without retropseudogenes, the former being, in their view, more widely expressed than the latter. However, in the human genome, using a different algorithm,
the propensity for retrotransposition was found to be unaffected by the GC content of the
genes (Venter et al., 2001). Studying the origin of CpG islands, Ponger et al. (2001) con-
firmed that tissue-specific genes were mainly localized in GC-rich isochores (64% in H3),
whereas the distribution of widely expressed genes was independent of the isochore con-
text.

It should be stressed that, in both papers dealing with the correlation between gene
distribution and expression patterns (Gonçalves et al., 2000; Ponger et al., 2001), the
gene partition was performed according to the criteria defined in Mouchiroud et al.
(1991). In the last decade, however, several results based on theoretical and experimental
approaches led to an improvement of the gene partition criteria (Saccone et al., 1993, 1999;
Zoubak et al., 1996; Federico et al., 2000). Using this partition criteria and two indepen-
dent datasets, the analysis of the gene frequencies in GC-poor and GC-rich regions led to
the conclusions that tissue-specific and widely expressed genes: 1) show no compositional
differences at GC3 level; and 2) follow the general gene distribution, reaching the highest
frequency in the GC-rich isochores (D'Onofrio, 2002; see Fig. 5.4). Needless to say, these
conclusions are in complete contradiction with those of Gonçalves et al. (2000) and Ponger
et al. (2001; see also note added in proof in Part 12, Chapter 6).

(iii) **Chromatin structure.** It has been known for a long time that the open chromatin
structure is characterized by accessibility to DNases (Kerem et al., 1984) and that CpG
island chromatin is characterized by a larger spacing of nucleosomes, absence of histone
H1 and acetylation of histones H3 and H4 (Tazi and Bird, 1990). Since CpG island density
parallels gene density (Aissani and Bernardi, 1991a,b; Jabbari and Bernardi, 1998; see Fig.
5.11), one should conclude that chromatin structure is largely open in the genome core, and
largely closed in the empty quarter. In fact, it is quite possible that open chromatin regions
are more abundant, even in absolute amounts, in the scarce GC-rich, gene-rich isochores than
in the abundant GC-poor, gene-poor isochores.

![Figure 5.4. Distribution of A widely expressed (housekeeping) genes and B tissue-specific genes. The three GC3 classes were the result of averaging the boundaries defined by theoretical (Zoubak et al., 1996) and experimental results (Saccone et al., 1993, 1999; Federico et al., 2000). (From D'Onofrio, 2002).](image-url)
(iv) **Recombination** has been shown to be more frequent in GC-rich compared to GC-poor isochores (see Nachman, 2002 for a review, and Part 7, Chapter 4).

(v) **Post-transcriptional control of gene expression** (see Fig. 5.5). 1) The efficiency of AUG start codon recognition in translation initiation is modulated by its sequence context. This context is different in genes located in different isochores. In particular, of the two main consensus start sequences, RCCatatg R is five-fold more represented than AARatatg R in genes from the GC-rich H3 isochores compared to genes from the GC-poor L isochore. Furthermore, genes located in GC-rich isochores have shorter 5' UTRs and stronger avoidance of upstream ATG compared to genes located in GC-poor isochores (Pesole et al., 1999). This suggests that genes requiring highly efficient translation are located in GC-rich isochores and genes requiring fine modulation of expression are located in GC-poor isochores. Interestingly, the most represented initiator context in mammalian genes, as well as in warm-blooded vertebrates genes in general, is GCCatatg G, whereas in cold-blooded vertebrates it is AAAtatg G (a context very similar to that observed in GC-poor human isochores), as expected for genomes in which GC-rich isochores are very scarce or absent (Pesole et al., 1999). 2) Both 3' and 5' UTR (untranslated regions), but especially the former, are shorter as their GC level increases, in parallel with GC3 (Pesole et al., 1997, 2000). It was also observed that the average GC of 5'UTRs is higher than that of 3'UTRs, this difference being more marked in warm-blooded vertebrates, 60% GC for 5'UTRs vs. 45% for 3'UTRs, probably because of the frequent presence of CpG islands (Pesole et al. 2001).
To sum up, an ample body of data has conclusively demonstrated that the genomes of mammals (and, more generally, of warm-blooded vertebrates; see the following section) are not only structural mosaics, made up of compositional compartments, the isochores, but also functional mosaics.

1.3. The distribution of genes in the vertebrate genomes

The obvious question as to how widespread among vertebrates is the distribution of genes exhibited by the human genome was approached experimentally in several ways. (i) Hybridization of human DNA corresponding to H3 isochores (in an excess of unlabeled Alu sequences in order to avoid hybridization of repeated sequences) on compositional fractions from other vertebrate genomes. Single-copy human DNA derived from H3 isochores hybridized on the GC-richest compositional fractions representing 10-15% of all mammalian and avian genomes tested, indicating a similar compositional distribution of the highest concentration of genes (Cacciò et al., 1994; Fig. 5.6). More interestingly, hybridization of human single-copy sequences from H3 isochores was also found in the 10-15% GC-richest compositional DNA fractions of cold-blooded vertebrates (see Fig. 5.6). (ii) Prediction of ORFs (open reading frames) in long DNA sequences. In *Brachidanio rerio*, the estimate of ORF density per 10 kb is 1.7 (sd 0.99) for 895 kb of GC-poor (35.1% GC) regions, and 3.5 (sd 1.82) for 617 kb of GC-rich (39.7% GC) regions, showing again an increased gene density in the GC-rich regions. (iii) Extending the approach of Fig. 5.1 to

![Figure 5.6. Slot blot hybridisation patterns obtained using as a probe the gene-richest fraction of human DNA on compositional fractions from vertebrate genomes. Species names are indicated at the top. The modal buoyant densities of the fractions increase from the top to the bottom, as indicated at the left of the figure, the last fractions often containing, however, satellite DNAs. Fraction numbers are indicated at the left of the figure. The amounts of DNA in the fractions are variable. In all cases, however, hybridization-positive fractions represent 10-15% of DNA. (Modified from Cacciò et al., 1994).](image-url)
the draft sequence of the *Fugu* genome and using the data of Fig. 4.5 allowed us to demonstrate a higher gene density in the GC-rich regions (Fig. 5.7).

One should, therefore, conclude that a **bimodal gene distribution is present in all vertebrates**, before and after the compositional transition separating cold- from warm-blooded vertebrates. This important conclusion is not completely surprising because the alternative hypothesis of vertebrate genes clustering in the newly formed GC-rich isochores of warm-blooded vertebrates would have implied a tremendous reshuffling of the genome. This conclusion is also supported by the growing evidence for the existence of syntenic regions not only among warm-blooded vertebrates, but also between species as phylogenetically distant as fishes and human.

The different functional properties of the two gene spaces were, therefore, present well before the two independent compositional transitions that took place in the reptilian ancestors of present-day mammals and birds and concerned only one of the two gene spaces, namely the ancestral genome core. Cases of changes in isochore location of some genes are also known, however, a specially interesting case being that of duplicated genes (see **Part 6**).
The distribution of CpG islands in the vertebrate genome

Unmethylated CpG islands (which will be referred to as CpG islands henceforth) are sequences 0.5-2 kb in size that are rich in GC and in unmethylated CpG doublets and that were originally investigated in the mouse genome using a restriction enzyme (Hpa II), which cuts the sequence CCGG producing Hpa II tiny fragments, or HTF (Bird et al., 1985), only if the C of the central CG dinucleotide is not methylated. CpG islands were reported in the genomes of mammals, birds, reptiles, amphibians and fishes (Cooper et al., 1983) and were found to be associated with the 5'-flanking sequences, 5' exons and 5' introns of all housekeeping genes tested so far, and of some tissue-specific genes, as well as with the 3' exons of some tissue-specific genes (Gardiner-Garden and Frommer, 1987).

When we investigated the distribution of CpG islands in vertebrate genomes fractionated according to their base composition (Aissani and Bernardi, 1991a,b), we found that genomes from warm-blooded vertebrates (man, mouse and chicken) are characterized by abundant CpG islands, whose frequency increases in DNA fractions of increasing GC, in parallel with the distribution of genes and CpG doublets (Fig. 5.8). Small, yet significant, differences in the distribution of CpG islands were found in the three genomes (see below). In contrast, genomes from cold-blooded vertebrates (two reptiles, one amphibian, and two fishes) were characterized, contrary to an earlier report (Cooper et al., 1983), by an extreme scarcity or absence of CpG islands (Figs. 5.9, 5.10; see Aissani and Bernardi, 1991a).

CpG islands associated with homologous genes from cold- and warm-blooded vertebrates were then compared by analyzing CpG frequencies, GC levels, Hpa II sites, rare-cutter sites and G/C boxes (GGGGCGGCGGC and closely related motifs) in sequences available in gene banks. Small, yet significant, differences were again detected among the CpG islands associated with homologous genes from warm-blooded vertebrates, in that CpG islands associated with mouse or rat genes often showed low CpG and/or GC levels, as well as low numbers of Hpa II sites, rare-cutter sites and G/C boxes, compared to homologous human genes; more rarely, CpG islands were just absent (Aissani and Bernardi, 1991a, b). This phenomenon, later called CpG island erosion (Matsuo et al., 1993), will be discussed in Part 11, Chapter 4. As far as cold-blooded vertebrates were concerned, a number of genes showed sequences, that exhibited a much lower frequency of CpG doublets than CpG islands of warm-blooded vertebrates, but still approached the statistically expected frequency. However, none of the other features of CpG islands associated with genes from warm-blooded vertebrates were present. Other genes did not show any associated CpG islands, unlike their homologs from warm-blooded vertebrates. The findings just outlined indicate that the origin and evolution of CpG islands in the vertebrate genome are associated with the compositional transitions (GC increases) in genes and isochores, which took place independantly in the ancestors of mammals and birds.

A computer analysis of human DNA sequences larger than 50 kb and representing about 120 Mb of DNA (Jabbari and Bernardi, 1998) led to the following observations:
Figure 5.8. CsCl profiles of compositional DNA fractions from A human placenta, B mouse liver and C chicken erythrocyte, as obtained from preparative centrifugation in Cs$_2$SO$_4$/BAMD density gradient. The $r_f$ (ligand/nucleotide molar ratio) values used were 0.12 for mouse and 0.14 for human and chicken. Modal buoyant densities and relative amounts of the fractions are indicated. P indicates the pellet. Notice the satellite peak (centered at about 1.700 g/ml) in the last fractions of human, and the satellite peak (1.691-1.692 g/cm$^3$) in mouse DNA fractions 1-4. $A'$, $B'$ and $C'$ display autoradiograms of terminally labelled (Cooper et al., 1983; 1.2% agarose gels were used) $Hpa$ II fragments from DNA fractions. (From Aissanm and Bernardi, 1991a).

(i) The frequency of CpG islands in long human sequences increases with increasing GC and almost parallels gene frequency (Fig. 5.11). (ii) GC and CpG levels of CpG islands are positively correlated with the GC levels of the long sequences in which they are located. (iii) CpG levels of CpG islands increase with their own GC levels.
Figure 5.9. CsCl profiles of compositional DNA fractions from A crocodile (*Crocodylus niloticus*) and B turtle (*Testudo graeca*) erythrocytes, as obtained from preparative centrifugation in Cs$_2$SO$_4$/BAMD density gradient. The *r*$_f$ values used in the fractionations of crocodile and turtle DNA were 0.14 and 0.16, respectively. Modal buoyant densities and relative amounts of the fractions are indicated. P, pellet A' and B' show autoradiograms of terminally labelled *Hpa* II fragments from DNA fractions. (See also legend of Fig. 5.8; from Aissani and Bernardi, 1991a).
Figure 5.10. CsCl profiles and *Hpa* II fragments of compositional DNA fractions from Xenopus, trout and carp DNAs. CsCl profiles of A *Xenopus laevis*, B trout (*Onchorhynchus mykiss*), and C carp (*Cyprinus carpio*) liver, as obtained from preparative centrifugation in Cs₂SO₄/BAMD density gradient (Bernardi and Bernardi, 1990a): *r* values of 0.14 were used. Modal buoyant densities and relative amounts are indicated. P, pellet. Satellite peaks or shoulders are visible in the CsCl profiles of fraction 1 of *Xenopus* DNA, fractions 6-9 of trout DNA and fractions 7 and 8 of carp DNA. Panels A', B' and C' show autoradiograms of *Hpa* II fragments from DNA fractions. In panel C', the pellet was very poorly labelled. (See also legend of Fig. 5.8; from Aissani and Bernardi, 1991a).

Figure 5.11. Density of A genes and B CpG islands in isochore families. Relative numbers of sequences over relative amounts of isochore families are presented in the histograms. (From Jabbari and Bernardi, 1998).
CHAPTER 3

The distribution of CpG doublets and methylation in the vertebrate genome

3.1. CpG doublets

It is well known that vertebrate DNAs show a characteristically low frequency of CpG doublets, whereas all other doublets are present at about the frequency expected from a random distribution of nucleotides (Josse et al., 1961; Swartz et al., 1962). Two explanations were proposed to account for this outstanding feature of the vertebrate genome, the so-called CpG shortage.

(i) The first explanation (Subak-Sharpe et al., 1966) was based on the observation that the CpG shortage is also exhibited by the genomes of small vertebrate viruses (like polyoma and SV40), which use essentially all their genetic information for directing protein synthesis. It was proposed (Subak-Sharpe et al., 1967, 1974) that the genomes of these viruses derived from polypeptide-specifying DNA of the host cells and, therefore, exhibited the same CpG shortage, which was visualized as reflecting constraints from the translation apparatus. In agreement with this proposal, CpG shortage was absent in tRNA, 5sRNA and rRNA genes. The lack of CpG shortage in the genome of intermediate and large viruses (such as adenovirus and herpes simplex virus, respectively) was attributed to their capacity to modify the host translation apparatus. Since only a very small proportion of vertebrate DNA is actively involved in protein coding and since a CpG shortage was found in all compositional fractions of guinea pig DNA, as obtained by gradient centrifugation, by differential renaturation or from dispersed and condensed chromatin, it was further concluded that "the bulk of vertebrate DNA derives from, and maintains the gross sequence characteristic of polypeptide-specifying DNA" (Russell et al., 1976).

This hypothesis was already somewhat compromised by the discovery that α globin mRNAs from rabbit (Salser, 1977) and human (Forget et al., 1979) carry a large number of CpG doublets, and by the different levels of CpG in both coding and non-coding sequences from the human α and β globin gene clusters (Lennon and Fraser, 1983). Subsequent investigations (Bernardi et al., 1985b; Bernardi, 1985; Bernardi and Bernardi, 1986a) definitely demonstrated that CpG shortage is not related to polypeptide coding. In fact, CpG shortage (i) is strong in GC-poor genes, but becomes increasingly weaker in GC-rich genes from warm-blooded vertebrates; and, moreover, (ii) decreases with increasing GC in non-coding (intrinsic and intergenic) sequences. In other words, CpG levels in both coding and non-coding sequences from warm-blooded vertebrates are correlated with the GC levels of the corresponding isochores. These findings not only contradicted the explanation proposed by Subak-Sharpe and co-workers for the CpG shortage, but also demonstrated that CpG shortage is not uniform throughout the genome, at least in the case of warm-blooded vertebrates, in contrast with the conclusion of Russell et al. (1976). Finally, as far as the different CpG levels exhibited by the genomes of small and large
vertebrates viruses are concerned, it was shown that these differences were not related to genome size, but simply to the GC levels of the corresponding genomes. Indeed, a plot of CpG vs GC for viral genomes exhibited a positive linear relationship (Bernardi and Bernardi, 1986a) which was very close (Fig. 5.12) to that previously found for vertebrate genes (Bernardi et al., 1985b).

(ii) An alternative explanation for the CpG shortage in vertebrates was based on the consideration that the C residues in CpG doublets are highly methylated in vertebrates (Sinsheimer, 1955; Doskocil and Sorm, 1962; Grippo et al., 1968; Van der Ploeg and Flavell, 1980; Gruenbaum et al., 1981; Kunnath and Locker, 1982), to the extent of 50-90% in different mammalian DNAs (Gruenbaum et al., 1981). It was proposed that mCpG represents a hot-spot for mutation (Salser, 1977; Coulondre et al., 1978) since it can be deaminated to TpG. Even in the presence of a repair system that specifically corrects mispaired T/G back to C/G (Kramer et al., 1984), the evolutionary effect will be a progressive loss of CpG dinucleotides, leading to a CpG shortage, and to a concomitant increase in TpG (and of its complementary doublet CpA on the other strand). Indeed, a scenario which was put forward for the evolution of DNA methylation in vertebrates (Cooper and Krawczak, 1989) is that the vertebrate genome originally was strongly methy-
lated (and GC-rich) and that the methylated CpG doublets (as well as the GC levels) subsequently underwent a monotonous decay. This explanation as well as the alternative hypothesis of an equilibrium between CpG depletion and CpG formation (Sved and Bird, 1990), which also predicts a continuous decrease in CpG during vertebrate evolution, were, however, also shown to be wrong by investigations from our laboratory (see Sections 3.2 and 3.3).

3.2. Two different CpG levels in vertebrate genomes

Table 5.2 displays the CpG frequency and the observed/expected ratios for the CpG dinucleotides, as well as the genomic GC levels. The data show that the genomes of fishes and amphibians ranging from Chondrychthyes to Anura have a similar CpG o/e (observed/expected) ratio, averaging 0.37, which is significantly higher than that, averaging 0.26, found for nine genomes from eight mammals and one bird. The fact that mammals from six orders share similar CpG o/e values and the star-like phylogeny of mammals (see Fig. 11.3) suggest that this value was also that of their last common ancestor (see Part 11, Section 3.3 for additional comments). Unfortunately, no data are available for reptiles.

<table>
<thead>
<tr>
<th>Genomic Species</th>
<th>GC</th>
<th>CpG</th>
<th>CpG o/e</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Squale acanthius</td>
<td>45.3</td>
<td>1.8</td>
<td>0.35</td>
<td>4</td>
</tr>
<tr>
<td>Salmo salar</td>
<td>43.1</td>
<td>1.7</td>
<td>0.37</td>
<td>1</td>
</tr>
<tr>
<td>Latimeria sp.</td>
<td>41.7</td>
<td>1.8</td>
<td>0.41</td>
<td>4</td>
</tr>
<tr>
<td>Rana catesbeiana</td>
<td>45</td>
<td>1.7</td>
<td>0.33</td>
<td>5</td>
</tr>
<tr>
<td>average</td>
<td>43.78</td>
<td>1.75</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>Oryctolagus cuniculus</td>
<td>44.3</td>
<td>1.3</td>
<td>0.25</td>
<td>1</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>43.9</td>
<td>1.2</td>
<td>0.25</td>
<td>2</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>42.2</td>
<td>1</td>
<td>0.23</td>
<td>1</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>44</td>
<td>1.4</td>
<td>0.29</td>
<td>1</td>
</tr>
<tr>
<td>Phocoena phocoena</td>
<td>41.4</td>
<td>1.3</td>
<td>0.33</td>
<td>3</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>42</td>
<td>1</td>
<td>0.25</td>
<td>1</td>
</tr>
<tr>
<td>Halichoerus sp.</td>
<td>39</td>
<td>0.9</td>
<td>0.24</td>
<td>3</td>
</tr>
<tr>
<td>Cavia porcellus</td>
<td>40</td>
<td>1</td>
<td>0.26</td>
<td>3</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>44.2</td>
<td>1.1</td>
<td>0.23</td>
<td>1</td>
</tr>
<tr>
<td>average</td>
<td>42.33</td>
<td>1.13</td>
<td>0.26</td>
<td></td>
</tr>
</tbody>
</table>

From Jabbari et al., 1997

References. (1) Swartz et al. (1962); (2) Skalka et al. (1966); (3) Russell (1974); (4) Russell and Subak-Sharpe (1977); (5) McGeoch (1970).

Bos taurus was excluded from the average because of large amounts (23%) of GC-rich satellites (Macaya et al., 1978).
3.3. Two different methylation levels in vertebrate genomes

Investigations aimed at understanding the biological significance of DNA methylation in vertebrates have been carried out along two lines.

(i) Studies on the correlation between promoter methylation and gene expression led to a number of interesting insights. Siegfried and Cedar (1997) recognized three different strategies by which methylation could affect gene expression: (1) methylation of CpG may interfere directly with the binding of specific transcription factors to DNA; this does not appear to affect gene-specific transcription factors, but rather ubiquitous factors; (2) the direct binding of specific factors to methylated DNA may lead to gene repression; (3) methylation may cause repression by altering chromatin structure; indeed, DNA methylation affects positioning of nucleosomes and influences the sensitivity of DNA to DNAses.

(ii) In contrast, analyses of overall DNA methylation levels in DNA from different vertebrate and different tissues (Vanyushin et al., 1970, 1973; Pollock et al., 1978; Ehrlich et al., 1982; Gama-Sosa et al., 1983a,b; Serrano et al., 1993) did not lead to any conclusion. A re-examination of the general problem of DNA methylation in vertebrates has, however, led to new insights.

Table 5.3 presents the GC and 5mC levels of the 42 vertebrate genomes studied by Jabbari et al. (1997), as well as those reported in the literature for 45 additional species. c values (haploid genome sizes) and the tissues from which DNAs were extracted are also indicated. It should be noted that, even in the sample of Table 5.3, the largest studied so far, vertebrate classes are not represented in a balanced way, fishes and mammals being, by far, predominant.

A summary of the results of Table 5.3 indicates (Table 5.4) that the average methylation values in fishes (1.70%) and amphibians (1.98%) are roughly twice as large as those found in mammals (0.88%), birds (1.02%) and reptiles (1.00%). Since average GC levels are close to 42.5% for all five vertebrate classes, this means that approximately 9% of all cytosines are methylated in fishes and amphibians compared with approximately 4.5% in mammals, birds and reptiles.

Fig. 5.13 shows that two statistically significant correlations ($p<0.0001$) hold between 5mC and genomic GC levels, one for fishes and amphibians and another one for mammals and birds. It should be noted that (i) a few mammalian values deviate from the correlation shown by mammals/birds; deviating points concern, beside the bovine somatic cells mentioned above, *Erinaceus europeus*, *Meriones unguiculatus*, *Cricetus norvegicus*, *Manis* sp. (lower values) and *Pteropus poliocephalus* (higher value), these deviations being possibly due to under- or over-methylated satellites, respectively; (ii) reptilian values show a larger scatter compared with values from mammals/birds (see Table 5.4 and Fig. 5.13) to which they are close; (iii) one of the fishes, the only chondrichthyan studied, showed an especially low value compared with other fishes, 1.15%.

The correlation between 5mC and genomic GC levels is understandable if one recalls that CpG frequency is linearly correlated with the GC content of the genome (Bernardi, 1985; Aissani and Bernardi, 1991a,b) and that CpG is the major site of methylation, methylation on other dinucleotides (Woodcock et al., 1987) and on CpNpG trinucleotides (Clark et al., 1995) being extremely limited. A similar correlation was also found in plant
TABLE 5.3.
5mC, GC and c-values in vertebrate DNAs.

<table>
<thead>
<tr>
<th>Order</th>
<th>Family</th>
<th>Species</th>
<th>Tissue</th>
<th>GC (mol%)</th>
<th>5mC (mol.%)</th>
<th>5mC ref.</th>
<th>c-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fishes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cyclostomata</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Petromyzontiformes</td>
<td>Petromyzontidae</td>
<td>Lampropterus flaviatilis</td>
<td>blood</td>
<td>50.9</td>
<td>2.09</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>Osteichthyes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acipenseriformes</td>
<td>Acipenseridae</td>
<td>Acipenser gueldenstadtii</td>
<td>sperm</td>
<td>43.8</td>
<td>1.79</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Huso huso</td>
<td>sperm</td>
<td>43.4</td>
<td>1.81</td>
<td>1</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Acipenser ruthenus</td>
<td>sperm</td>
<td>43.0</td>
<td>1.87</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Acipenser stellatus</td>
<td>sperm</td>
<td>42.9</td>
<td>1.87</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Antheriniformes</td>
<td>Oryzias latipes</td>
<td>liver</td>
<td>40.8</td>
<td>1.85</td>
<td>5</td>
<td>1</td>
<td>1.1</td>
</tr>
<tr>
<td>Clupeiformes</td>
<td>Clupeidae</td>
<td>Clupea harengus</td>
<td>sperm</td>
<td>45.2</td>
<td>1.87</td>
<td>1</td>
<td>0.77</td>
</tr>
<tr>
<td>Cypriniformes</td>
<td>Cyprinidae</td>
<td>Cyprinus carpio</td>
<td>sperm</td>
<td>38.9</td>
<td>1.56</td>
<td>1</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Brachydanio rerio</td>
<td>liver</td>
<td>37.3</td>
<td>1.27</td>
<td>1</td>
<td>1.8</td>
</tr>
<tr>
<td>Siluriformes</td>
<td>Siluridae</td>
<td>Silurus glanis</td>
<td>liver</td>
<td>41.1</td>
<td>1.64</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ictaluridae</td>
<td>Ictalurus nebulosus</td>
<td>liver</td>
<td>41.3</td>
<td>1.92</td>
<td>5</td>
<td>1.2</td>
</tr>
<tr>
<td>Salmoniformes</td>
<td>Esocidae</td>
<td>Esox lucius</td>
<td>sperm</td>
<td>43.5</td>
<td>1.87</td>
<td>1</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sperm</td>
<td>44.2</td>
<td>1.98</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sperm</td>
<td>43.5</td>
<td>1.47</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Salmonidae</td>
<td>Osmerus eperlanus</td>
<td>Sperm</td>
<td>45.4</td>
<td>2.01</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oncorhynchus sp.</td>
<td>sperm</td>
<td>44.0</td>
<td>1.64</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>liver</td>
<td>44.1</td>
<td>2.13</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oncorhynchus tschawitscha</td>
<td>sperm</td>
<td>41.9</td>
<td>1.73</td>
<td>1</td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Oncorhynchus mykiss</td>
<td>liver</td>
<td>42.9</td>
<td>1.45</td>
<td>5</td>
<td>2.8</td>
<td></td>
</tr>
<tr>
<td>Gadiformes</td>
<td>Gadidae</td>
<td>Gadus morhua</td>
<td>blood</td>
<td>50.0</td>
<td>2.24</td>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Eleginus narvaga</td>
<td>blood</td>
<td>48.2</td>
<td>2.15</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Scorpaeniformes</td>
<td>Cottidae</td>
<td>Cottus sp.</td>
<td>sperm</td>
<td>41.0</td>
<td>1.70</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
TABLE 5.3.
5mC, GC and c-values in vertebrate DNAs\textsuperscript{b} (continued)

<table>
<thead>
<tr>
<th>Order</th>
<th>Family</th>
<th>Species</th>
<th>Tissue</th>
<th>GC (mol%)</th>
<th>5mC (mol.%)</th>
<th>5mC ref.\textsuperscript{b}</th>
<th>c-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perciformes</td>
<td>Scombridae</td>
<td>Trichiurus japonicus</td>
<td>sperm</td>
<td>41.3</td>
<td>1.50</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thynnidae</td>
<td>Thunnus thynnus</td>
<td>sperm</td>
<td>37.0</td>
<td>1.54</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Cichlidae</td>
<td>Oreochromis grahami</td>
<td>liver</td>
<td>41.5</td>
<td>1.29</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oreochromis niloticus</td>
<td>liver</td>
<td>41.0</td>
<td>1.26</td>
<td>6</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tilapia buttikofleri</td>
<td>liver</td>
<td>40.3</td>
<td>1.46</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mugilidae</td>
<td>Mugil cephalus</td>
<td>sperm</td>
<td>39.5</td>
<td>1.87</td>
<td>2</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>liver</td>
<td>39.5</td>
<td>1.80</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Pleuronectiformes</td>
<td>Pleuronectidae</td>
<td>Pleuronectes flesus</td>
<td>blood</td>
<td>44.9</td>
<td>1.70</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Tetrodonsiformes</td>
<td>Tetraodontidae</td>
<td>Liopsetta glacialis</td>
<td>blood</td>
<td>45.2</td>
<td>1.84</td>
<td>1</td>
<td>0.73</td>
</tr>
<tr>
<td>Amphibia</td>
<td>Leptodactylidae</td>
<td>Odontophrynus americanus</td>
<td>blood</td>
<td>41.9</td>
<td>1.63</td>
<td>6</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Pipidae</td>
<td>Xenopus laevis</td>
<td>liver</td>
<td>42.2</td>
<td>1.40</td>
<td>1</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>Ranidae</td>
<td>Rana temporaria</td>
<td>liver</td>
<td>45.1</td>
<td>2.60</td>
<td>1</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>blood</td>
<td>45.1</td>
<td>2.61</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rana pipiens</td>
<td>liver</td>
<td>47.3</td>
<td>2.30</td>
<td>1</td>
<td>5.8</td>
</tr>
<tr>
<td>Reptilia</td>
<td>Crocodylia</td>
<td>Crocodylidae</td>
<td>Crocodylus niloticus</td>
<td>blood</td>
<td>44.4</td>
<td>1.06</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Chelonia</td>
<td>Testudinidae</td>
<td>Testudo horsfieldi</td>
<td>blood</td>
<td>45.9</td>
<td>1.30</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>liver</td>
<td>44.8</td>
<td>1.50</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Squamata</td>
<td>Varanidae</td>
<td>Varaonis griseus</td>
<td>blood</td>
<td>43.7</td>
<td>1.15</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Ophidida</td>
<td>Crotalidae</td>
<td>Bothrops neuriodi</td>
<td>blood</td>
<td>40.0</td>
<td>1.01</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bothrops jararac</td>
<td>blood</td>
<td>39.7</td>
<td>0.97</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Boa constrictor amarali</td>
<td>blood</td>
<td>40.4</td>
<td>0.61</td>
<td>6</td>
<td>3.15</td>
</tr>
<tr>
<td>Aves</td>
<td>Rheiformes</td>
<td>Rheidae</td>
<td>Rhea americana</td>
<td>blood</td>
<td>44.0</td>
<td>0.98</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Sphenisciformes</td>
<td>Spheniscidae</td>
<td>Spheniscus demersus</td>
<td>blood</td>
<td>43.3</td>
<td>1.01</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Anseriformes</td>
<td>Anatidae</td>
<td>Cairina moschata</td>
<td>blood</td>
<td>41.5</td>
<td>0.64</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Galliformes</td>
<td>Phasianidae</td>
<td>Gallus gallus</td>
<td>blood</td>
<td>44.2</td>
<td>1.04</td>
<td>1</td>
</tr>
</tbody>
</table>
TABLE 5.3.
5mC, GC and c-values in vertebrate DNAs. (continued)

<table>
<thead>
<tr>
<th>Order</th>
<th>Family</th>
<th>Species</th>
<th>Tissue</th>
<th>GC (mol%)</th>
<th>5mC (mol.%)</th>
<th>5mC ref.(^{b})</th>
<th>c-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Columbiformes</td>
<td>Columbidae</td>
<td>Columba livia</td>
<td>blood</td>
<td>44.0</td>
<td>1.02</td>
<td>6</td>
<td>1.3</td>
</tr>
<tr>
<td>Passeriformes</td>
<td>Corvidae</td>
<td>Corvus corone</td>
<td>blood</td>
<td>45.4</td>
<td>1.15</td>
<td>1</td>
<td>1.7</td>
</tr>
<tr>
<td>Ratitiformes</td>
<td>Rallidae</td>
<td>Fulica atra</td>
<td>blood</td>
<td>49.3</td>
<td>1.33</td>
<td>1</td>
<td>1.7</td>
</tr>
<tr>
<td>Mammalia</td>
<td>Monotremata</td>
<td>Ornithorhynchidae</td>
<td>Ornithorhynchus anatinus</td>
<td>blood</td>
<td>45.6</td>
<td>1.32</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Tachyglossidae</td>
<td>Tachyglossus aculeatus</td>
<td>blood</td>
<td>45.9</td>
<td>0.99</td>
<td>6</td>
<td>3.57</td>
</tr>
<tr>
<td>Eutheria</td>
<td>Insectivora</td>
<td>Erinaceidae</td>
<td>Erinaceus europaeus</td>
<td>thymus</td>
<td>45.5</td>
<td>0.53</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Soricidae</td>
<td>Crocidura russula</td>
<td>liver</td>
<td>41.4</td>
<td>0.66</td>
<td>6</td>
<td>—</td>
</tr>
<tr>
<td>Dermoptera</td>
<td>Cynocephalidae</td>
<td>Cynocephalus variegatus</td>
<td>liver</td>
<td>40.6</td>
<td>0.89</td>
<td>6</td>
<td>—</td>
</tr>
<tr>
<td>Megachiroptra</td>
<td>Pteropodidae</td>
<td>Pteropus sp.</td>
<td>liver</td>
<td>40.5</td>
<td>0.96</td>
<td>6</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Pteropodidae</td>
<td>Pteropus poliocephalus</td>
<td>liver</td>
<td>41.5</td>
<td>1.39</td>
<td>6</td>
<td>—</td>
</tr>
<tr>
<td>Microchiroptra</td>
<td>Hipposideridae</td>
<td>Hipposideros galeritus</td>
<td>whole animal</td>
<td>41.4</td>
<td>0.86</td>
<td>6</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Rhinolophidae</td>
<td>Rhinolophus creughi</td>
<td>whole animal</td>
<td>41.5</td>
<td>0.81</td>
<td>6</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Vespertilionidae</td>
<td>Myotis lucifugus</td>
<td>whole animal</td>
<td>43.5</td>
<td>0.97</td>
<td>6</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>Phyllostomidae</td>
<td>Chiroderma solvinii</td>
<td>liver</td>
<td>41.4</td>
<td>0.86</td>
<td>6</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Nycteridae</td>
<td>Nycteris thebaica</td>
<td>whole animal</td>
<td>42.9</td>
<td>0.88</td>
<td>6</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Molossidae</td>
<td>Chaerephon pumila</td>
<td>liver</td>
<td>41.4</td>
<td>0.86</td>
<td>6</td>
<td>2.9</td>
</tr>
<tr>
<td>Primates</td>
<td>Hominidae</td>
<td>Homo sapiens</td>
<td>sperm</td>
<td>42.6</td>
<td>0.84</td>
<td>3</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>blood</td>
<td>42.6</td>
<td>0.96</td>
<td>3</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Cercopithecidae</td>
<td>Macaca mulatta</td>
<td>liver</td>
<td>41.3</td>
<td>0.86</td>
<td>4</td>
<td>3.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Macaca fascicularis</td>
<td>liver</td>
<td>41.3</td>
<td>0.96</td>
<td>4</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cercopithecus aethiops</td>
<td>liver</td>
<td>41.3</td>
<td>0.98</td>
<td>4</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Cebidae</td>
<td>Saimiri sciureus</td>
<td>liver</td>
<td>41.3</td>
<td>0.85</td>
<td>4</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Tupaiidae</td>
<td>Tupai montana</td>
<td>liver</td>
<td>41.9</td>
<td>1.08</td>
<td>6</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Lemuridae</td>
<td>Hapalemur griseus</td>
<td>liver</td>
<td>41.4</td>
<td>0.93</td>
<td>6</td>
<td>—</td>
</tr>
<tr>
<td>Pholidota</td>
<td>Monidae</td>
<td>Mans sp.</td>
<td>liver</td>
<td>42.4</td>
<td>0.59</td>
<td>6</td>
<td>—</td>
</tr>
<tr>
<td>Lagomorpha</td>
<td>Leporidae</td>
<td>Oryctolagus cuniculus</td>
<td>liver</td>
<td>44.3</td>
<td>0.86</td>
<td>6</td>
<td>3.2</td>
</tr>
<tr>
<td>Rodentia</td>
<td>Sciuridae</td>
<td>Sciurus vulgaris</td>
<td>liver</td>
<td>39.5</td>
<td>0.61</td>
<td>6</td>
<td>5.1</td>
</tr>
<tr>
<td>Order</td>
<td>Family</td>
<td>Species</td>
<td>Tissue</td>
<td>GC (mol%)</td>
<td>5mC (mol.%)</td>
<td>5mC ref.&lt;sup&gt;b&lt;/sup&gt;</td>
<td>c-value</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td>----------------------------</td>
<td>--------</td>
<td>-----------</td>
<td>-------------</td>
<td>------------------------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td>Cricetidae</td>
<td><em>Cricetus norvegicus</em></td>
<td>liver</td>
<td>40.7</td>
<td>0.37</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Spalacidae</td>
<td><em>Spalax sp.</em></td>
<td>liver</td>
<td>38.4</td>
<td>0.58</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Caviidae</td>
<td><em>Cavia porcellus</em></td>
<td>liver</td>
<td>39.7</td>
<td>0.74</td>
<td>6</td>
<td>2.95</td>
</tr>
<tr>
<td></td>
<td>Muridae</td>
<td><em>Rattus norvegicus</em></td>
<td>liver</td>
<td>43.9</td>
<td>0.94</td>
<td>6</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Mus musculus</em></td>
<td>liver</td>
<td>42.2</td>
<td>0.95</td>
<td>4</td>
<td>3.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sperm</td>
<td>42.2</td>
<td>0.83</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Cetacea</td>
<td>Procaviidae</td>
<td><em>Procavia capensis</em></td>
<td>liver</td>
<td>41.0</td>
<td>0.70</td>
<td>6</td>
<td>3.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Meriones unguiculatus</em></td>
<td>nd</td>
<td>45.2</td>
<td>0.70</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Carnivora</td>
<td>Canidae</td>
<td><em>Canis familiaris</em></td>
<td>liver</td>
<td>41.1</td>
<td>0.67</td>
<td>6</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>Felidae</td>
<td><em>Panthera uncia</em></td>
<td>liver</td>
<td>41.5</td>
<td>0.94</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Perissodactyla</td>
<td>Equidae</td>
<td><em>Equus caballus</em></td>
<td>liver</td>
<td>42.8</td>
<td>1.04</td>
<td>6</td>
<td>2.95</td>
</tr>
<tr>
<td>Arniocactyla</td>
<td>Bovidae</td>
<td><em>Bos taurus</em></td>
<td>liver</td>
<td>44.0</td>
<td>1.40</td>
<td>4</td>
<td>3.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>kidney</td>
<td>44.0</td>
<td>1.30</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sperm</td>
<td>44.2</td>
<td>0.75</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Suidae</td>
<td><em>Sus scrofa</em></td>
<td></td>
<td>kidney</td>
<td>44.0</td>
<td>1.20</td>
<td>4</td>
<td>2.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sperm</td>
<td>42.1</td>
<td>0.77</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Ovidae</td>
<td><em>Ovis aries</em></td>
<td></td>
<td>sperm</td>
<td>42.0</td>
<td>1.13</td>
<td>4</td>
<td>2.9</td>
</tr>
</tbody>
</table>

<sup>a</sup> (From Jabbari et al., 1997)

<sup>b</sup> 5mC values are from (1) Vanyushin et al. (1970, 1973); (2) Pollock et al. (1978), Gama-Sosa et al. (1983b); (3) Ehrlich et al. (1982); (4) Gama-Sosa et al. (1983a); (5) Serrano et al. (1993) and (6) present work. GC values are from Bernardi and Bernardi (1990a) for fishes, amphibians and reptiles, Kadi et al. (1993) for birds and Sabour et al. (1993) for mammals. c-values are from literature quoted in the above papers and from Bachmann (1972) and Venturini et al. (1986)
TABLE 5.4.
5mC and GC levels in different classes of vertebrates

<table>
<thead>
<tr>
<th>Class</th>
<th>Species</th>
<th>GC%</th>
<th>sd(^b)</th>
<th>5mC</th>
<th>sd(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fishes(^c)</td>
<td>28</td>
<td>42.44</td>
<td>2.87</td>
<td>1.70</td>
<td>0.26</td>
</tr>
<tr>
<td>Amphibians</td>
<td>4</td>
<td>44.13</td>
<td>2.56</td>
<td>1.98</td>
<td>0.56</td>
</tr>
<tr>
<td>Reptiles</td>
<td>7</td>
<td>42.75</td>
<td>2.63</td>
<td>1.00</td>
<td>0.26</td>
</tr>
<tr>
<td>Birds</td>
<td>7</td>
<td>44.53</td>
<td>2.41</td>
<td>1.02</td>
<td>0.21</td>
</tr>
<tr>
<td>Mammals</td>
<td>39</td>
<td>42.07</td>
<td>1.68</td>
<td>0.88</td>
<td>0.20</td>
</tr>
<tr>
<td>Vertebrates</td>
<td>85</td>
<td>42.55</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) From Jabbari et al. (1997)
\(^b\) sd, standard deviation
\(^c\) Only Osteichthyes were taken into consideration.
Figure 5.13. Plots of 5mC levels against GC levels for the genomes of fishes/amphibians (solid circles), reptiles (open squares), mammals (open circles) and birds (open triangles). When data from different tissues were available, mean values were used. Deviating points of mammals are in grey (two of them, indicated by arrows, corresponding to DNAs from bovine somatic and germ cells, respectively) and were not taken into account in drawing the regression lines of mammals and birds; the single deviating point from fishes is shown as a solid triangle. Correlation coefficients, slopes and intercepts are given for fishes/amphibians and mammals/birds, respectively. (From Jabbari et al., 1997).

Figure 5.14. A. Plot of 5mC against the percentage of repetitive DNA sequences of the corresponding genomes (from Ginatuln, 1984). Solid circles correspond to fishes/amphibians, open circles to mammals/birds. B. Plot of 5mC levels against the corresponding haploid genome size of vertebrates; data from Table 5.3 were used to construct the plot. C. GC levels of fish DNAs are plotted against the corresponding haploid genome sizes; data are from Table 3 of Bernardi and Bernardi (1990a). e-values were used only up to 2pg, higher values being due to polyploidization. In all cases the correlation coefficients, slopes, intercept and p-values are shown. (From Jabbari et al., 1997).
genomes (Matassi et al., 1992; see Part 8) and in isochore families from individual vertebrate genomes (human, mouse, chicken and *Xenopus*; see below).

The two different methylation levels in vertebrate genome might be related to the different relative amounts of repetitive DNA sequences since (i) repetitive DNA sequences, as isolated by reassociation experiments, are more methylated than single-copy DNA sequences, at least in mammals (Romanov and Vanyushin, 1981; Gama-Sosa et al., 1983b); and (ii) the proportion of repetitive DNA sequences is larger in fishes/amphibians (59-77%) than in reptiles/mammals/birds (27-41%) (see Olmo et al., 1989, for a review). As expected from these premises, when the relative amount of repetitive DNA sequences (from Ginatulin, 1984) was plotted against the methylation level of the corresponding genomes, a positive, statistically significant correlation was found (see Fig. 5.14A).

There are, however, some problems with the relationship of Fig. 5.14A. (i) While the correlation persisted when plotting only the results from five fishes and one amphibian, a non-significant correlation was found when plotting only data from birds and mammals. (ii) At least one of the fish points of Fig. 5.14A is very doubtful. Indeed, *Clupea harengus*, with a c value of only 0.77 pg (see Table 5.3), was estimated to comprise 62% repetitive sequences, leaving only 0.29 pg for non-repetitive DNA. Now, the work of Pizon et al. (1984) has shown that in *Arothron diadematus*, one of the Tetraodontid fishes that have the smallest genome size of all vertebrates, 0.45 pg, repetitive sequences amount to 13%, leaving 0.39 pg for non-repetitive DNA. This result casts a serious doubt on the *C. harengus* results, and decreases the points supporting the correlation to only five. More importantly, the results of Pizon et al. (1984) led to the conclusion that the reduction in genome size, such as occurred in *Tetraodontiformes*, largely took place at the expense of repetitive sequences. Conversely, this conclusion confirms the suggestion (Olmo et al., 1989) that genome size increases not due to polyploidization largely take place by an increase of repetitive sequences. (3) The genomes of fishes characterized by the lowest c values and by the lowest amount of repetitive sequences (*Tetraodontiformes* and *Gadidae*) show a methylation level that is similar to those of fishes with a genome size at least twice as large. Along the same lines, there is no substantial difference in methylation level between the two classes of warm-blooded vertebrates, in spite of the fact that birds have a genome size close to one-half that of mammals and a much larger amount of slow-reassociating DNA (84% in chicken DNA vs. 62% in human DNA; Soriano et al., 1981; Olofsson and Bernardi, 1983). It should, therefore, be concluded that the difference in methylation level between fishes/amphibians and reptiles/mammals/birds is not correlated with the amounts of repetitive sequences within either one of the two sets of genomes.

This conclusion is confirmed by investigations concerning the correlation between methylation and c values, the haploid genome sizes. As discussed above, c values are related to the amounts of repetitive sequences. In fact, they have the advantage of providing a larger data set compared with reassociation data. As shown in Fig. 5.14B, a negative correlation holds between 5mC levels and c values. This negative correlation can be understood on the basis of the following considerations. A negative correlation holds between GC and genome size in the case of fish genomes (Bernardi and Bernardi, 1990a; see Fig. 5.14C). This indicates that genome expansion (in a c value range where polyploidization is absent) takes place by an increase in the size of intergenic regions, which, as a general rule (Clay et al., 1996), are GC-poorer than coding sequences and, as a consequence, have fewer
methylatable sites. Therefore, increasing the genome size means increasing GC-poorer regions of the genome which leads to a relative decrease of methylation, so accounting for the results of Fig. 5.14B.

5-Methylcytosine (5mC) levels were also determined in compositional DNA fractions corresponding to different isochore families from the genomes of *Xenopus*, chicken, mouse and human, four vertebrates which show different isochore patterns (Cacciò et al., 1997). The results obtained indicate that: (i) positive correlations exist between the 5mC levels and the GC levels of isochores within any given genome (Fig. 5.15); (ii) DNA from *Xenopus* isochore families is twice as methylated as DNA from the isochores having the same GC levels from mouse, human and chicken; (iii) higher methylation levels are associated with higher GC levels of main-band fractions; in some cases plots display deviating points that correspond to hypermethylated fractions (usually satellite DNAs). A similar result was obtained for plant genomes by Matassi et al. (1992; see Fig. 8.18).

In conclusion, our findings completely contradicted the explanations for CpG shortage proposed by Subak-Sharpe et al (1966) and Russell et al (1976), as well as those proposed by Cooper and Krawczak (1989) and Sved and Bird (1990). An alternative explanation based on our data on two different CpG and methylation levels in cold- and warm-blooded vertebrates will be presented in Part 11, Section 3.5.

![Figure 5.15](image_url)

Figure 5.15. Plot of 5mC vs GC for compositional DNA fractions from *Xenopus*, chicken, mouse and man (left-hand scale). The histograms indicate 5mC levels. Least square lines through the points (and their equations) are shown; only filled circles were used in drawing the lines. The *Xenopus* (broken) line is also shown in the other diagrams for comparison. Likewise, the mouse and chicken lines are shown in the human diagram. (From Cacciò et al., 1997)
Part 6

The distribution of integrated viral sequences, transposons and duplicated genes in the mammalian genome
The distribution of proviruses in the mammalian genome

1.1. The integration of retroviral sequences into the mammalian genome

The integration of retroviral genomes as proviruses into the genomes of host cells is a critical step in the life cycle of retroviruses (Temin, 1976; Varmus, 1984). Indeed, the replication of retroviruses is dependent upon integration (see Goff, 1992; Brown, 1997, for reviews), and several pathogenic effects of retroviruses are associated with their integration. For example, integration may inactivate host cell genes by disruption, activate them by the action of viral promoters and enhancers, or modulate their activity (reviewed in Kung et al., 1991; Tsichlis and Lazo, 1991; Athas et al., 1994; Jankers and Berns, 1996). Integration may be also associated with the modification of cellular genes and with the formation of new transforming viruses by acquisition of cellular gene sequences (reviewed in Bishop, 1980; Neil, 1983; Hughes, 1983). These events may lead to cell transformation and cancer development. It is evident, therefore, that retroviral integration is a very important phenomenon. Retroviral integration is, in fact, a special case of integration of foreign DNA into the genome of host cells. Obviously, a good understanding of this case has a crucial importance in gene therapy, especially when retroviruses are used as vectors (reviewed in Bushman, 1995; Hodgson, 1996; Anderson, 1998).

Integration of retroviral genomes (see Fig. 6.1) is a site-specific process as far as the viral sequences are concerned, since long terminal repeat (LTR) sequences are involved (reviewed in Varmus and Brown, 1989; Grandgenett and Mumm, 1990). As for the host genome, local features of integration loci (nucleosomal versus non-nucleosomal DNA, bent versus unbent DNA etc.), as well as the role of integrase in target site selection (reviewed in Holmes-Son et al., 2001), have been investigated, but while "it is not clear to what extent local effects can account for the pattern of integration over the whole genome", it has been concluded that "the distribution of integration sites across the whole genome remains to be adequately characterized" (Brown, 1997). Investigations carried out in our laboratory since the late 1970's, taking into account the isochore organization of the genome of warm-blooded vertebrates, have led, however, to an understanding of this distribution.

1.2. The bimodal compositional distribution of retroviral genomes

Compositionally, retroviral genomes belong in two classes (Zoubak et al., 1992; Tsyba et al., 2003): a GC-poor class and a GC-rich class (Fig. 6.2A,B). The first, minor, class comprises all lentiviruses and spumaviruses (which have no oncogenes but contain genes for regulatory proteins), oncoviruses of B type and some oncoviruses of D type, like MMTV (which also do not contain oncogenes). The second, major, class includes all
oncoviruses containing oncogenes (like RSV) and some oncoviruses that do not contain oncogenes (like BLV and HTLV-1), except for those of B type and for some oncoviruses of the D type. The striking bimodal compositional distribution of retroviral genomes is accompanied by a similar bimodal distribution of LTRs, which are GC-rich or GC-poor according to the GC level of the viral sequences, and by a remarkable compositional homogeneity of genes (gag, pol and env genes, as well as oncogenes and genes for regulatory proteins) within each retroviral genome (Zoubak et al., 1992).

1.3. The localization of integrated viral sequences in the host genome

The localization of integrated proviruses in the isochores of the host genome can be approached in exactly the same way as the localization of genes described in the preceding
section. In the case of proviruses, restriction enzymes that do not cut proviral sequences are typically used, so that the number of different-size bands reflects the number of proviral copies. Hybridization patterns are generally characterized by different distributions of proviruses in the compositional fractions, different spreads of same-size bands and different intensities of the bands. Fig. 6.3 shows two examples of proviral localization. It should be stressed that the hybridization signals allow detecting of even the viral sequences that originate from small sub-populations of cells that diverged from the main population because of rearrangements of cellular and/or viral DNA. These proviral sequences appear as faint hybridization bands, which are expectedly distinct from the faint bands corresponding to the tails of distribution of same-size bands.

The compositional localization of integrated retroviral sequences in the mammalian genome was investigated for three GC-rich viruses, bovine leukemia virus, BLV (Kettmann et al., 1979; 1980), Rous sarcoma virus, RSV (Rynditch et al., 1991), human T-cell leukemia virus 1, HTLV-1 (Zoubak et al., 1994), and for two GC-poor viruses, mouse mammary tumor virus, MMTV (Salinas et al., 1987) and human immunodeficiency virus 1, HIV-1 (Glukhova et al., 1999). Some common features of the localization of integrated proviruses have been observed, such as compartmentalization, isopycnicity (see below), and correlation between transcription of the proviral sequences and isopycnicity. These features will be illustrated by describing the results obtained with different viral systems.

BLV. A compartmentalized localization of integrated BLV proviral sequences (Fig. 6.4) was demonstrated in the genomes of infected cells obtained from three animals, one with persistent lymphocytosis and two in the tumor stage. Proviral copies were found at a large number of genomic sites in the leukocytes of the animals with persistent lymphocytosis and at one or very few sites in the genomes of leukocytes or tumor cells of animals in the tumor stage of the disease. In all six samples investigated (from leukocytes, lymph node tumors and spleen), BLV sequences were only found in DNA fragments present in GC-rich compositional fractions (>49% GC), which represent much less than 10% of the bovine genome, if one neglects the presence in the same buoyant density range of several satellite DNAs (that make up 23% of the bovine genome; see Part 3) and of rDNA (Filipski et al., 1973; Macaya et al., 1978; Meunier-Rotival et al., 1979). As BLV has a GC-rich genome (54% GC), such localization into the GC-rich compartments of the host genome was defined as an “isopycnic localization”, meaning that integrated sequences are found in isochores approximately matching the viral genome in buoyant density and base composition.

Figure 6.3. Scheme of the localization of HTLV-1 proviral sequences in the compositional fractions and total DNA from the genome of a human cell line A and a cell clone B, as obtained after centrifugation in a preparative CsSO₄/BAMD density gradient. In each pattern, the first line presents the fraction number, the second line the GC level. Relative intensities of hybridization bands are indicated by lines of three different thicknesses. (Modified from Zoubak et al., 1994)
Figure 6.4. A. Distribution of proviruses (BLV, RSV, MMTV) in the genome of calf (Kettmann et al., 1979), hamster (Rynditch et al., 1991) and mouse (Salinas et al., 1987). The CsCl profiles of host cell DNAs (with buoyant densities converted into GC levels) are shown along with the localization of proviral sequences (open bars). Arrows indicate the GC levels of the viral genomes. B. Distribution of integrated viral sequences in the human genome: HTLV-1 (Zoubak et al., 1994) and HBV (Zerial et al., 1986b). The profile of human DNA is shown by the solid line. Arrows indicate the GC level of the viral sequences. (From Rynditch et al., 1998).

**RSV.** In this case, the localization of integrated retroviruses was investigated in the genomes of six well-characterized clones of hamster tumor cells. Twelve out of thirteen proviral sequences were observed in DNA fragments present in compositional fractions corresponding to the GC-richest 20% of the hamster genome, with a peak at 50% GC (Fig. 6.4). More precisely, GC-rich RSV (54% GC) from the four cell lines derived from different tumors containing one or two complete or defective copy(s) were found in DNA fractions having GC levels equal to 50–53% (Rynditch et al., 1991). The direct sequencing of extended DNA stretches flanking the provirus in one such cell line confirmed its integration
in GC-rich DNA regions (Machon et al., 1996). In the case of cell lines comprising both complete and defective proviral copies, complete copies were present in fractions centered at about 50% GC, whereas defective copies were found in these fractions as well as in GC-poorer ones.

**HTLV-I.** The localization of this GC-rich provirus (54% GC) was investigated (Fig. 6.4) in five immortalized cell lines, containing a total of 22 integrated complete or defective proviruses, and in seven T-cell clones obtained from patients with TSP/HAM (tropical spastic paraparesis/HTLV-1-associated myelopathy), in which case each clone comprised 1–3 integrated proviral sequences for a total of 18 proviruses (Zoubak et al., 1994). These forty proviruses were found in the 39–54% GC range of the human genome, whereas no HTLV-I sequences were found in isochores having a GC level lower than 39% (in spite of these isochores representing as much as 40% of the human genome).

**MMTV.** GC-poor MMTV proviral sequences (43.3% GC) were localized in the DNAs from the livers of five inbred strains of mice and from GR cells derived from primary implants of mammary tumors (Salinas et al., 1986). Twelve sets of endogenous MMTV sequences (corresponding to 27 proviruses) were characterized by restriction patterns and chromosomal localizations. Exogenous sequences were also localized and produced strong hybridization signals (the total number was 13), or weak hybridization signals corresponding to proviral integrations in a minor cell population (the total number was 6). While endogenous sequences were present in the GC-poor 43% of the host genome, exogenous sequences showed a broader distribution, being present in the GC-poor 60% of the genome (Fig. 6.4), therefore still within the L isochores. Interestingly, endogenous sequences were centered at 40% GC, whereas exogenous sequences showed a broader distribution centered at 43% GC.

To sum up, four retroviruses, three of C-type (BLV, RSV, HTLV-1) and one of D-type (MMTV), were localized in four mammalian species. The data show that integrated retroviral sequences are not spread at random in the host genomes, but are found in some host genome compartments approximately matching the viral sequences in composition. This conclusion is stressed by a presentation of the results in terms of proviral density in host DNA (Fig. 6.5). Expectedly, in none of the cases investigated were the proviral sequences found in satellite or ribosomal DNA.

Results similar to those just described for retroviral sequences were obtained for the localization of integrated GC-rich sequences from human hepatitis B virus (HBV), a DNA virus (Zerial et al., 1986b). Eight out of the nine viral sequences integrated in the hepatoma Alexander cell line were localized in DNA segments having a GC level of 51% (Fig. 6.4). Likewise, integration of a GC-rich recombinant SV40-adenovirus 5 was observed in the H3 isochores family (Romani et al., 1993).

It should be stressed that our conclusions are not contradicted by results indicating, for instance, that “HTLV-I integration is not random at the level of the nucleotide sequence. The virus was found to integrate in A/T rich regions” (Leclercq et al., 2000), because these investigations dealt, in fact, with a different problem, concerning the short (40-nucleotide long) host sequences flanking the provirus. This very local composition may obviously be present in any isochore. It should also be stressed that our results concern the localization of stably integrated proviral sequences and not the primary events of integration (which will be discussed below).
1.4. An analysis of integration sites near host cell genes

Another approach that we used to localize integrated retroviral sequences should be mentioned. Since host cell genes are "isochore markers", in that their composition is correlated with that of the isochores in which they are embedded, an analysis of proviruses located near host cell genes can be done in order to localize proviruses (Rynditch et al., 1998). This was done for 64 MuLV (Murine Leukemia Virus) sequences for which information about the GC level of several genes in the neighborhood of integration sites was available. This analysis of host genes located close to MuLV proviruses showed that these genes are distributed in a GC range centered around 54% GC (Fig. 6.6), a value very close to the GC level of MuLV (53%). This means that in the case of insertional mutagenesis, it is possible to observe an isopycnic localization of integrated proviral sequences.
1.5. Correlation between the isochore localization of integrated retroviral sequences and their transcription

The results on the compositional distribution of HBV sequences (Zerial et al., 1986b) provided the first information about a possible correlation between the isopycnic localization of integrated viral sequences in compositional compartments of the host genome and their transcription. Indeed, the eight HBV sequences integrated in the GC-richest regions of the genome of the Alexander cell line were expressed, whereas the single one located in a GC-poorer region of the genome was not (see Fig. 6.4). Likewise, all RSV transcribed sequences were localized in the GC-richest isochores (Fig. 6.4), whereas the proviral sequences whose expression was not detected were distributed in isochores with a lower GC content (Rynditch et al., 1991).

A more detailed analysis was possible on integrated HTLV-1 sequences, in which case 40 sequences from 12 cell lines and T-cell clones were investigated (Zoubak et al., 1994). Indeed, the transcriptionally active HTLV-1 sequences were found to be localized in the GC-richest regions which corresponded to the H2 and H3 isochore families of the human genome, whereas the transcriptionally inactive HTLV-1 sequences were localized in GC-rich regions corresponding to the H1 isochore family (Fig. 6.4).

In the case of MMTV, at least two loci, MTV-1 and MTV-2, located in GC-poor isochores (37–38%) are expressed as infectious viruses in GR (Bentvelzen et al., 1970) and C3H (Van Nie and Verstaelen, 1975) strains of mice and in the GR mammary tumor cell line. Moreover, most mouse strains contain at least one copy of endogenous MMTV provirus (Marcus et al., 1981) expressed not only in mammary glands during lactation (Choi et al., 1987), but also in several organs at a lower level (Henrard and Ross, 1988). These data indicate that MMTV sequences can, indeed, be expressed when integrated in GC-poor isochores.
1.6. Integration in "open" chromatin and/or near CpG islands

Several observations suggested that retroviral localization is correlated with the level of transcription and recombination, and with the degree of DNA accessibility in chromatin (Vijaya et al., 1986; Robinson and Gagnon, 1986; Rohdewohld et al., 1987; Scherdin et al., 1990; Mooslehner et al., 1990; Fincham and Wyke, 1991).

In bursal lymphomas, ALV (Avian Leukemia Virus) integrants were found near each of five major DNase I hypersensitive sites which are immediately upstream of the coding sequences for c-myc (Schubach and Groudine, 1984; Robinson and Gagnon, 1986). The integration of MoMuLV (Moloney Murine Leukemia Virus) sequences in the α-1 collagen gene has been localized near a DNase I hypersensitive site (Breindl et al., 1984). Likewise, MoMuLV integrations in seven regions of chromosomal DNA were mapped near DNase hypersensitive sites (Vijaya et al., 1986). Three of these regions, containing c-erbB, c-myc and dsi-1, were targets for multiple tumor-inducing integrations. Interestingly, the other four unselected integrations also occurred near DNase I hypersensitive sites, suggesting that retroviral sequences preferentially integrate near hypersensitive sites in all cases. This observation has been confirmed by the analysis of chromatin structure of MoMuLV proviral integration sites in early embryonic cells and in differentiated fibroblasts, where integration occurs within a few hundred base pairs of a DNase I hypersensitive chromatin region (Rohdewohld et al., 1987). It is well known that DNase I hypersensitive sites correlate with gene expression and are preferentially located in regions with an open chromatin structure (Edmondson and Roth, 1996).

These results prompted the analysis of transcription of cellular sequences flanking the integrated proviruses. In three out of five randomly chosen mouse strains harboring one copy of MoMuLV in their germ lines, the provirus was integrated in the vicinity of DNA regions transcribed in an embryonal stem cell line and in an embryonal carcinoma cell line (Mooslehner et al., 1990). Among nine sequences randomly chosen for MoMuLV proviral integration sites in early embryonic cells and in differentiated fibroblasts, at least six were in transcriptionally active regions and/or contained CpG-rich islands (Scherdin et al., 1990). Finally, the proviral integration sites of transcribed and non-transcribed RSV in rat DNA were examined (Fincham and Wyke, 1991). In this case, the transcribed sequences showed a tendency to integrate close to the 3' ends of CpG islands, whereas non-transcribed sequences did not.

A more recent study (Carteau et al., 1998) of 61 HIV-1 integration site sequences after a short (14 hrs) experimental infection of a human T-cell line led, however, to the conclusion of no clear bias for integration in transcription units, but that centromeric alphoid repeats were absent at integration sites. An analysis of 48 such sites in terms of genome localization (Elleder et al., 2002) showed, however, that 54% of them were located in genes (introns) and 69% in transcribed sequences, in line with the older results mentioned above. As far as the isochores localization is concerned, as shown in Fig. 6.7, while the HIV-1 integrations are mainly found in isochores L2, H1 and H2, their highest density (and, therefore, their preference of integration) is in the genome core (isochores H2 and H3), which is characterized by an open chromatin structure (see Parts 4, 5 and 7). This is a most interesting result in view of the fact that HIV-1 has a very GC-poor genome.

The conclusions of Elleder et al. (2002) have been confirmed by a more extensive study on the sites of HIV-1 integration in the human genome (Schröder et al., 2002). Needless to
say, the gene dense-regions in which the preferential localization of these early integration sites occurs coincide (although apparently not noticed by the authors) with the GC-richest isochores (see Fig. 6.8). An interesting aspect of the work of Schröder et al. (2002) was that these authors not only investigated HIV-1 integration in vivo (as already mentioned), but also in vitro. Using naked genomic DNA from Sup T1 cells as an integration target and HIV-1 preintegration complexes (i.e. replication intermediates that can be isolated from infected cells and which contain the viral cDNA, integrase, and other viral and cellular proteins), they found an integration pattern which did not show the specificity for genes

Figure 6.7. A. Numbers of HIV-1 sequences in each isochore family, the GC intervals are L1 (<37%), L2 (37–41%), H1 (41–46%), H2 (46–52%) and H3 (>52%). The size of the genome belonging to individual isochore families in 100 kb windows is 571 Mb for L1, 977 Mb for L2, 706 Mb for H1, 321 Mb for H2 and 85 Mb for H3. B. The vertical axis shows the probability of targeting the individual isochore family. This is calculated as the number of integrations per Gb whose 100 kb surrounding belongs to the corresponding isochore family. (Modified from Elleder et al., 2002).

Figure 6.8. Distribution of 500 in vivo and 110 in vitro HIV-1 integration sites in the human genome based on the data from Schröder et al. (2002). The curves represent the integration density (% integrations/ % DNA). (From Tsyba et al., 2004).
shown in vivo. Plotting the results of Schröder et al. (2002) against GC levels of 100-kb genome regions, we found, however, that integration, although characterized by lower levels, mimicked the results obtained in vivo (Fig. 6.8). This important result suggests that the pre-integration complexes basically have a specificity for GC-rich DNA, in all likelihood in its protein-DNA complex form. In other words, the artificial chromatin complexes of GC-rich DNA seem to be a preferred target for integration.

Very interestingly, the localization found for such early integrations changes with time. An investigation on 27 HIV-1 patients (Tsyba et al., 2003) showed that the increasing production of virus-specific RNA copies is correlated with provirus localization. Representative samples show that low levels of RNA copies/ml (up to 4,000) is correlated with a predominance of proviruses in the GC-rich isochores (Fig. 6.9A), medium levels (5,000-16,000) with a localization in GC-poor isochores as important as that in GC-rich isochores (Fig. 6.9B) and high levels (60,000-280,000) again with a preferential localization in GC-rich isochores (Fig. 6.9C). We interpret these results as indicating that the production of virus-specific RNA copies increases from the first to the second period because transcription is increased for proviruses located in the compositionally matching GC-poor environment, whereas the late phase is the result of novel integrations from the abundant viral RNA produced in the intermediate period. Not surprisingly, in persistently infected cells, the provirus was found in GC-poor or GC-intermediate isochores (Glukhova et al., 1999).

In conclusion, the results available suggest that the initial integration into open chromatin regions (such as those corresponding to the GC-richest isochores) is favoured even for a GC-poor provirus. Stability of integration and expression are, however, associated with an isopycnic localization.

1.7. The causes of the compartmentalized, “isopycnic localization” of viral sequences

The results just presented suggest that (i) integration of proviral sequences is initially
targeted towards “open” chromatin regions; possibly because of their abundance (see Part 5), the “open” chromatin regions of GC-rich isochores are the main targets for integration of retroviral sequences independently of the GC level of the proviral sequence; as expected, no observation of integration of proviruses in heterochromatic regions, in which repeated sequences are characterized by a tightly closed chromatin structure, was reported so far; (ii) isopycnicity is associated with stability of integration; indeed, “open chromatin regions” exist in both GC-rich and GC-poor isochores, yet even non-expressed integrated retroviral sequences tend to show an isopycnic localization in the genome; along the same line, the majority of interspersed repeated sequences, which are not expressed, show their highest densities in isopycnic isochores (see Chapter 2); (iii) transcription of integrated viral sequences appears to be associated, as a rule, with an isopycnic localization, as indicated by several examples (HIV-1, HTLV-1 etc.) and by transcribed integrated sequences that act in \textit{trans} (see below); this is a very important point because it implies the existence of functional correlations between coding and noncoding sequences, leading to the concept of \textit{genomic fitness} (see Part 12); (iv) selection for a replicative advantage may play a role in the choice of specific sites within an isopycnic region; (v) in exceptional cases (such as MMTV activating GC-rich oncogenes) selection may override isopycnicity (see below).

Other considerations of interest here are the following:

(i) As far as oncogene activation is concerned, the target of proviral integration, namely the regions flanking the oncogene, is so small that it is most unlikely that selection for those integration sites operates on an initially completely random integration pattern. In contrast, if genome regions accessible to viral integration are limited to those undergoing active transcription, this obviously restricts greatly the potential integration sites and reduces strongly the load imposed on selection. Moreover, one would also expect that, since open chromatin regions are more abundant in GC-rich isochores, targeting would be primarily directed to them. (ii) A compartmentalized, isopycnic localization is also found for proviruses acting in \textit{trans} (such as BLV and HTLV-1). This indicates that selection for a growth advantage may be associated with the expression of genes located far away from the integration site. Since this gene expression is dependent upon the transcription of a provirus and the production of \textit{trans}-activating factors, the isopycnic location of the provirus indicates a dependence of its own transcription on a matching compositional environment. This dependence can be easily understood because it mimics that of host genes, which are compositionally correlated with their genomic environment, this correlation being the result of an adaptive process that took place over evolutionary times. (iii) If one considers the best known and most widespread case of GC-rich proviruses, like MuLV or ALV, acting in \textit{cis} and the fact that oncogenes are mostly GC-rich, one might think that the compartmentalized, isopycnic localization is just the result of selection for integration sites conferring a replicative advantage to the cell through the activation of a nearby oncogene. According to this interpretation, the GC-richness of both the viral sequences and of the oncogenes would be purely coincidental. Some exceptions to the “isopycnicity rule” seem to lend, however, further support to an explanation based on “selection”. Indeed, in rare cases, the provirus is not in an isopycnic chromosomal environment, but still is next to an oncogene which is activated by it (an example being the activation of GC-rich genes by MMTV), suggesting that selection for a replicative advantage can override
isopycnic integration. It should be noted, however, that the examples of non-isopycnic integration practically only concern a GC-poor provirus, MMTV, being integrated next to GC-rich genes, a point which is now better understood in view of the results on HIV-1 (see preceding section). Indeed, the opposite case, namely that of GC-rich proviruses activating GC-poor oncogenes, has only been found once (in the case of MuLV activating c-ki-ras), in spite of the fact that the number of GC-rich proviruses explored was much larger than that of GC-poor proviruses.

These conclusions have some implications concerning both the compositional evolution of retroviral genomes and gene therapy: (i) the bimodal compositional distribution of retroviral sequences may be visualized as the result of their compositional coevolution with the sequences of the host genome; it is conceivable that the compositional transitions, which led to the formation of GC-rich isochores of warm-blooded vertebrates, also led to similar compositional transitions (acting on their integrated forms) of what are now the GC-rich retroviral sequences (Zoubak et al., 1992); the results on duplicated genes reported in Chapter 3 reinforce this explanation; and (ii) the reasons for the poor expression of retroviral constructs used in gene therapy (Anderson, 1998) should be investigated not only at the level of the promotors present in the constructs, but also at the level of the integration sites which are used by the constructs; in many cases, constructs are characterized by compositional features which are never found in the host genome.
CHAPTER 2

The distribution of repeated sequences in the mammalian genome

2.1. Alu and LINE repeats in human isochores

Alus are short (~300 bp) GC-rich, non-autonomous elements, derived from 7SL RNA. They make up to 10% of the human genome (Lander et al., 2001). Mutational databases show that Alus may transpose into the introns of human genes (Schmid, 1988; Batzer and Deininger, 2002). Full-length LINEs are long (6–8 kb) GC-poor sequences encoding an RNA binding protein and a reverse transcriptase/endonuclease. LINE1 elements represent the most abundant group of LINEs (see Ostertag and Kazazian, 2001, for a review) and correspond to 17% of the human genome (see Fig. 6.10).

Compositional fractionation of human DNA not only allowed us to characterize the genome in terms of isochore families and gene densities, but also to investigate the distribution of interspersed repeated sequences. Reassociation kinetics provided the first indication that the distribution of intermediate repetitive sequences was different in DNA fractions from different isochore families (Soriano et al., 1981; Olofsson and Bernardi, 1983b). Indeed, as shown in Fig. 3.7, the relative amounts of interspersed repeats and unique sequences were strikingly different in the different major components of human, mouse and chicken genome. In a more direct approach, hybridisation of appropriate probes on the major DNA components (Fig. 6.11) showed that the GC-poor LINE-1 and the GC-rich Alu families have their highest densities in GC-poor and GC-rich isochores of the human and mouse genome, respectively (Meunier-Rotival et al., 1982; Soriano et al., 1983; Zerial et al., 1986b). These results, later confirmed by assessments based on sequences from data banks (Smit, 1996; 1999; Jabbari and Bernardi, 1998), showed that
the “ubiquitous” distribution of Alu sequences (Houck et al., 1979), was far from homogeneous.

The location of the long interspersed mouse LINE-1 repeats (originally called Bam H1 repeats) in the two light isochrome families L1 and L2 of mouse showed a compositional match with the isochromes in which they were embedded, and was strikingly different from that of protein-encoding genes, indicating that the repeats were not involved, at least in a simple way, in the regulation of gene expression (Meunier-Rotival et al., 1982). Subsequent work indicated not only that the human LINE-1 repeats (which were originally called Kpn I repeats) were also located in light isochromes, but also that the short interspersed SINE repeats (of the B1 family of mouse and of the Alu-1 family of human) were most frequent in the GC-rich isochromes. These results showed that the genome distribution of interspersed repeats is non-uniform, different for the two major families of repeats, and similar in two mammalian species. Moreover, the base composition of both classes of repeats was correlated with that of the isochromes in which they had their highest concentration, and both sorts of repeats were shown to be transcribed (Soriano et al., 1983).

Recent assessments have quantified the distribution of Alus and LINES in the human genome and provided information on the specific distribution of four categories of repeats separated according to their divergence, $\Delta$, from the consensus sequence (see Table 6.1). Very young Alus with $\Delta < 2\%$ (1,563 of the analyzed 280,809 Alus) were found to be twice as frequent in L1 compared to H3 isochromes. Alus with $\Delta = 2-4\%$ showed no preference, and Alus with $\Delta > 4\%$ have the typical bias toward the GC-rich part of the genome. The oldest and most abundant class of Alus $\Delta > 6\%$ have a more than threefold higher density in H3 compared to L1 isochore, and correspond to the general distribution of Alu sequences in the human genome.
Alus are CpG-rich and contain about 30% of all CpGs in the genome (Hellmann-Blumberg et al., 1993). Since the high rate of mutation of CpGs could affect age estimation based on the divergence from the consensus, Alu classification based on diagnostic sites was also used as another criterion (Batzer et al., 1996; Jurka, 2000; Fig. 6.12 presents a recent phylogenetic tree of Alu sequences). The young family AluYa5 (Deininger and Slagel, 1988) has a genomic distribution similar to the least divergent Alus (Fig. 6.13). The family AluYb8 is already more dense in GC-rich isochores, and AluY and the other families (members of AluS and AluJ group) are biased toward GC-rich isochores.

In the case of LINES, the preference for GC-poor isochores is much higher (90 times more in L1 than in H3 isochores) for young LINES of L1HS family, the only interspersed repeats that still actively transpose in the human genome (Lander et al., 2001), compared to old LINES (5 times), which practically show the general distribution (Fig. 6.14). The general distribution of both LINES and Alus are shown in Fig. 6.15.

We also calculated the Alu and LINE1 distribution on human chromosomes 21 and 22 (Dunham et al., 1999; Hattori et al., 2000). The densities of Alu and LINE1 co-vary with the GC level, LINES showing their highest density in GC-poor isochores and Alus being in GC-rich isochores. Insertions of the young Alu family AluYa5 seemed to be independent of the general Alu distribution on chromosomes 21 and 22 (Figs. 6.16A; 6.16B). The distribution of LINE L1PA2 family is strongly correlated with regions of low GC content. To test this point, the Alu (or LINE) density around each Alu (or LINE) in the database was analyzed within 20 kb flanking sequences. We found that the density of Alus surrounding each Alu insertion is higher for old families. Similar conclusions hold for LINES, older LINES being preferentially located within LINE clusters. We concluded that both Alus and LINES integrate both inside and outside clusters of older elements and that there are chromosomal regions that are more favorable for their accumulation (GC-poor for LINES

---

<table>
<thead>
<tr>
<th>DFC</th>
<th>L1</th>
<th>L2</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ &lt; 2%</td>
<td>0.077</td>
<td>0.081</td>
<td>0.068</td>
<td>0.046</td>
<td>0.033</td>
</tr>
<tr>
<td>Δ 2%-4%</td>
<td>0.062</td>
<td>0.052</td>
<td>0.069</td>
<td>0.073</td>
<td>0.069</td>
</tr>
<tr>
<td>Δ 4%-6%</td>
<td>0.2</td>
<td>0.22</td>
<td>0.36</td>
<td>0.47</td>
<td>0.56</td>
</tr>
<tr>
<td>Δ &gt; 6%</td>
<td>4.9</td>
<td>7.7</td>
<td>13.3</td>
<td>18.1</td>
<td>16.5</td>
</tr>
<tr>
<td>All Alus</td>
<td>5.2</td>
<td>8.1</td>
<td>13.8</td>
<td>18.7</td>
<td>17.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DFC</th>
<th>L1</th>
<th>L2</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ &lt; 2%</td>
<td>0.1008</td>
<td>0.0818</td>
<td>0.0410</td>
<td>0.0103</td>
<td>0.0011</td>
</tr>
<tr>
<td>Δ 2%-4%</td>
<td>1.41</td>
<td>1.27</td>
<td>0.67</td>
<td>0.17</td>
<td>0.012</td>
</tr>
<tr>
<td>Δ 4%-6%</td>
<td>1.1</td>
<td>1.2</td>
<td>0.48</td>
<td>0.11</td>
<td>0.031</td>
</tr>
<tr>
<td>Δ &gt; 6%</td>
<td>19.9</td>
<td>15.8</td>
<td>10.1</td>
<td>6.8</td>
<td>4.3</td>
</tr>
<tr>
<td>All LINES</td>
<td>22.5</td>
<td>18.3</td>
<td>11.3</td>
<td>7.1</td>
<td>4.4</td>
</tr>
</tbody>
</table>

*From Pavlček et al. (2001). Isochore family intervals are: less than 37%, 37% to 41%, 41% to 46%, 46% to 52%, and more than 52% GC. Repeat densities were calculated for 10 kb-long non overlapping sections using RepeatMasker. Alus and LINES are divided into four categories corresponding to less than 2%, 2-4%, 4-6% and more than 6% divergence (Δ) from the family consensus (DFC) sequence. The general Alu and LINE distributions are also shown (bold figures).
Figure 6.12. The expansion of Alu subfamilies (Yc1, Ya5a2, Yb9, Y, Sg1, Six and J) is superimposed on a
tree of primate evolution. The expansion of the various Alu subfamilies is colour coded to denote the time of
peak amplification. The approximate copy numbers of each Alu subfamily are also noted. Mya, million
years ago. (From Batzer and Deminger, 2002).

and GC-rich for Alus). Remarkable exceptions are the GC-rich telomeric regions of the
chromosomes, which are relatively free of Alus and rich in LINEs.

A final, important point is the following. Although the repeat density distribution is so
different for Alus and LINES (as shown for the general distribution in Fig. 6.15A), the
actual numbers of repeats follow rather similar trends (Fig. 6.15B). The largest number of
Alus and LINES are present in L2 and the lowest in H3 isochores, as expected from the
relative amounts of these isochores families. Evident differences exist in L1 isochores, where
LINES are more abundant than Alus and in H1-H3 isochores where the opposite is true.
Figure 6.13. Human genome distribution of Alu sequences in terms of numbers of sequences (A) and of density as calculated in 100 kb-long nonoverlapping segments (B). Isochore family intervals are as in Table 6.1. Distributions concern AluYa5 family (mean $\Delta = 2.26\%$); AluYb8 family (mean $\Delta = 5.23\%$); AluY family (mean $\Delta = 7.15\%$); AluJo family (mean $\Delta = 16.8\%$). (Updated from Pavliček et al., 2001).

Figure 6.14 Human genome distribution of LINE 1 sequences in terms of numbers of sequences (A) and of density as calculated in 100 kb-long non-overlapping segments. (From Pavliček, 2002, personal comm.)

Figure 6.15. General human genome distribution of Alu and LINE 1 sequences in terms of numbers of sequences (A) and of density as calculated in 100 kb-long non-overlapping segments. (From Pavliček, 2002, personal comm.)
2.2. The evolutionary origin of repeat distribution: different viewpoints

According to Lander et al. (2001), “the preference of LINEs for AT-rich DNA seems like a reasonable way for a genomic parasite to accommodate its host, by targeting gene-poor AT-rich DNA and thereby imposing a lower mutational burden. Mechanistically, selective targeting is nicely explained by the fact that the preferred cleavage site of the LINE endonuclease is TTTT/AA (where the slash indicates the point of cleavage), which is used to prime reverse transcription from the poly(A) tail of LINE RNA.”

As far as SINEs are concerned, the authors raise the question “How do SINEs accumulate in GC-rich DNA, particularly if they depend on the LINE transposition machinery (Jurka, 1997)?”. According to the authors, “One possibility is that SINEs somehow target GC-rich DNA for insertion. The alternative is that SINEs initially insert with the same proclivity for AT-rich DNA as LINEs, but that the distribution is subsequently reshaped by evolutionary forces (Smit, 1999; Arcot et al., 1998).” Having tacitly ruled out possibility 1 (see Table 6.2), and having shown that “recent Alus show a preference for AT-rich DNA resembling that of LINEs” (in spite of a difference by a factor of 45; see Table 6.1) they ask the question “What is the force that produces the great and rapid enrichment of Alus in GC-rich DNA?” and conclude that “this could be a higher rate of random loss of Alus in AT-rich DNA, negative selection against Alus in AT-rich DNA or positive selection in favour of Alus in GC-rich DNA. The first two possibilities (2a and 2b of Table 6.2) seem unlikely because AT-rich DNA is gene-poor and tolerates the accumulation of other transposable elements. The third seems more feasible, in that it involves selecting in favour of the minority that lie in GC-rich regions rather than against the majority that lie in AT-rich regions. But positive selection for Alus in GC-rich regions would imply that they benefit the organism.” The latter point was judged by the authors to be important enough to be included among the main conclusions of the paper. In line with this point, the authors propose “that Alu correlates not with GC content but with actively transcribed genes”. Moreover, they suggest that repeated sequences may “modulate” the overall GC content.

Unfortunately, these proposals go against well established facts. (i) Concerning the modulation of composition by repeated sequences, this possibility was already ruled out by old experiments (see Fig. 3.10) showing that GC variation was less than 1% on denatured and reassociated DNA over a C_0.1 range of 4 logs (Soriano et al., 1981). In other words, in spite of very large differences in repeat concentration, the composition of total DNA and of its compositional fractions was not different in slow, intermediate and fast reassociated fractions. Indeed, isochore composition is not remarkably different whether one includes or excludes repeated sequences (Pavliček et al., 2001; Fig. 6.17). In fact, after elimination of repeats, GC-rich sequences become slightly GC-richer and GC-poor sequences become slightly GC-poorer (Pavliček et al., 2001), two trends opposite to those expected from the putative “modulation” (Fig. 6.17).

(ii) “The idea that Alu correlates not with GC content but with actively transcribed genes” is contradicted by the finding that the highest density of Alu elements is in H2 isochores (Zerial et al., 1986a; Jabbari and Bernardi, 1998; Pavliček et al., 2001), whereas the highest density of genes is in H3 isochores (Zerial et al., 1986a; Mouchiroud et al., 1991; Zoubak et al., 1996). Moreover, the most gene-rich and GC-rich telomeric regions of chromosomes 21 and 22 are relatively free of Alus (see Fig. 6.16A).
A

GC profile

GC level (%)

Density (%)

LINE distribution

Density (%)

B

GC profile

GC level (%)

Density (%)

Alu distribution

Density (%)

LINE distribution

Density (%)

Figure 6.17 Correlation between the GC level of 100kb-long segments before and after exclusion of all repeats. (From Pavliček et al., 2001).

(iii) The hypothesis of positive selection favoured by Lander et al. (2001) has at least two major problems. The first one is of a general nature, namely that, however important, positive selection operates very rarely in evolution (see Brookfield, 2001). The second one is the lack of evidence for any advantage on which positive selection could work. The hypothesis of Schmid (1998), that they mention, postulates that the transcription of some SINEs under conditions of stress produces RNAs that specifically bind a particular protein kinase (PKR) which blocks the ability of PKR to inhibit protein translation. Such a promotion of protein translation under stress, while very interesting in itself, should however concern the majority of Alus located in GC-rich isochores in order to account for the proposal, whereas the hypothesis of Schmid (1988) involves a small number of Alus and does not say anything about their isochore localization.

(iv) As already mentioned, the possibility that Alus target GC-rich DNA (possibility 1 of Table 6.2) is ruled out on no ground at all.

(v) Finally, possibility 2 of Table 6.2 is not taken into account "because GC-poor DNA is gene-poor and tolerates the accumulation of other transposable elements", a reasoning which neglects the problems associated with the presence of sequences that do not compositionally match with the chromosomal environment of the host.

Instead of a highly problematic positive selection, we propose an alternative explanation, which takes into account some known facts: (i) the different frequency of acceptor
TABLE 6.2
Explanations for the accumulation of SINEs in GC-rich isochores discussed by Lander et al. (2001)

1. SINEs target GC-rich isochores for insertion
2. Young SINEs target GC-poor isochores for insertion but
   a) there is a higher rate of random loss of SINEs from GC-poor isochores
   b) SINEs are eliminated from GC-poor isochores by negative selection
3. SINEs are positively selected in GC-rich isochores

Modified from Bernardi (2001).

sites in GC-poor and GC-rich isochores; (ii) the preferential integration of retroviruses into open chromatin and their subsequent stability and expression in a compositionally matching chromosomal environment (see Chapter 1); and (iii) the different distribution of SINES and LINES, both as young or old elements.

Although there is a difference in the frequency of acceptor sites TTTT/AA over the 30% GC range of isochores, this difference might be compensated by the accessibility of sites which are located in mostly closed chromatin structures in the case of GC-poor isochores and in mostly open chromatin structures in the case of GC-rich isochores. We consider unlikely that the different frequency of acceptor sites plays a major role as the cause of the different distribution of SINES and LINES. Indeed, this difference seems to provide only a very slight advantage to the preferential integration of young Alus into GC-poor isochores, as opposed to a large advantage for the preferential integration of LINES into GC-poor isochores. This is a strong indication that other factors play a compensatory role in an unequal way in the case of Alus and in that of LINES. These factors are likely to be the compositional match between the transposon and the chromosomal region of the host, and the degree of interference of the integrated sequence with the function of neighboring genes.

In fact, because of a good compositional match, Alu sequences are not selected against when integrated into GC-rich isochores (as they are in GC-poor isochores), and accumulate there, changing their density ratio between GC-rich and GC-poor isochores. In contrast, LINES are not tolerated in GC-rich isochores, where their composition and size would alter the composition of GC-rich isochores and affect gene expression. Obviously, LINES have a good compositional match with GC-poor isochores, where they may accumulate. A point which needs to be stressed is that while stability is possibly comparable after an isopycnic integration of repeats, elimination is not. The small size of Alus, compared to LINES, and the vastity of GC-poor intergenic sequences make their integration into GC-poor isochores less deleterious and negative selection of the mismatched regions less rapid. In contrast, the large size of LINEs and reduced size of GC-rich intergenic sequences make this compositionally “mismatched” integration more deleterious and negative selection more rapid. A schematic presentation of the above explanations is given in Table 6.3. Needless to say that the obvious selective control on transposon distribution in the genome is not in agreement with the concept of selfish DNA that was applied to such sequences (Doolittle and Sapienza, 1980; Orgel and Crick, 1980).
TABLE 6.3
Proposed explanations for the densities of Alus and LINEs in GC-rich and GC-poor isochores

<table>
<thead>
<tr>
<th>Integration into:</th>
<th>Favored by:</th>
<th>Compositional Consequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alus:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GC-rich isochores</td>
<td>more open chromatin</td>
<td>+</td>
</tr>
<tr>
<td>GC-poor isochores</td>
<td>more acceptor sites</td>
<td>–</td>
</tr>
<tr>
<td><strong>LINES:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GC-rich isochores</td>
<td>more open chromatin</td>
<td>–</td>
</tr>
<tr>
<td>GC-poor isochores</td>
<td>more acceptor sites</td>
<td>+</td>
</tr>
</tbody>
</table>

2.3. Repeated sequences in coding sequences?

The contribution of transposable elements (TEs), including Alus, to human coding sequences has recently been reported to be high, 4% (1.3% Alus) out of 13,799 sequences (Nekrutenko and Li, 2001; Li et al., 2001). This is surprising, because previous examinations had revealed only very few repeats, and almost no Alus, in coding sequences (Brosius, 1999; Smit, 1999, Lander et al., 2001). Since extreme caution about input data has been suggested (Tugendreich et al., 1994; Zietkiewicz et al., 1994), we examined the database of Nekrutenko and Li (2001) and found that many (~30%) of its TE-containing sequences or their protein products are defined as ‘hypothetical’, and 63% (421/669 sequences) are annotated as ‘predicted, without experimental evidence or records without final NCBI revision’. Such a dataset is likely to contain several sequences that remain untranscribed, and more that remain untranslated. Not even experimental validation (Iyer et al., 2001), let alone computer prediction of functional genes, is foolproof: the errors in coding sequence databases such as those used in Nekrutenko and Li (2001) may well amount to 1-2% or more.

Essentially all reported coding regions derived from Alus, or containing alternatively spliced Alus, have been detected at the RNA (cDNA) level, instead of at the protein level (Brosius, 1999; Hoenika et al., 2002). In eukaryotic cells, there is a significant turnover of RNA, and several steps of quality control exist for the synthesized RNA in both nucleus and cytoplasm (Jackson et al., 2000; Wilusz et al., 2001; Maquat and Carmichael, 2001; Moore, 2002; Iborra et al., 2001). mRNAs with an aberrant 3'P end are generally retained and/or degraded at their site of transcription (Hilleren et al., 2001) and the majority of stable RNA polymerase II transcripts remain in the nucleus as ‘junk’ RNA, so they never reach the cytoplasm (Jackson et al., 2000). The minority of transcripts that are successfully exported from the nucleus undergo additional check(s) during their translation. For example, there are specialized degradation mechanisms for transcripts having premature stop codons or lacking terminal codons, which prevent the creation of aberrant, potentially pathogenic proteins (Wilusz et al., 2001; Moore, 2002; Frischmeyer et al., 2002). Thus, even detection of a transcript at the mRNA (cDNA) level cannot guarantee that these mRNAs are ever translated into stable proteins. As has been summarized in the light of growing evidence (Pradet-Balade et al., 2001), ‘mRNA abundance is a poor indicator of the
levels of the corresponding protein’, yet ‘it is the proteome that determines cell phenotype’: the transcriptome does not faithfully represent the proteome. Furthermore, to become a viable protein, a transcript must (after its accurate translation and possible post-translational modification) resist degradation until it can serve its functional role at the site of its required action. These facts underline the importance of detection at the protein level, for elucidating whether SINEs or other repeats contribute to true coding sequences in humans or mice.

The most accurate sources of proteins are 3D structure databases and direct amino acid sequencing. Out of 781 non-redundant human proteins from a 3D database or determined at the amino acid level and compared to human repeats in RepBase (Jurka, 2000) using TFASTX (Pearson et al., 1997), we found no Alu-related protein domain (the best hit has an E-value of 0.5). Twenty-eight apparently significant hits with E-values under 0.01 were detected, but mainly from protein-coding elements (DNA transposons and LINE1). When cDNAs encoding these 28 proteins were extracted and searched by RepeatMasker (Smit and Green), no interspersed repeats were detected. In addition, the similarity regions that had been reported by TFASTX were also found in other vertebrate orthologs. In summary, we did not detect any repeat sequence in our dataset of 781 protein sequences.

In 1994, it was pointed out (Tugendreich et al., 1994) that a discovery of a translated Alu element(s) in a functional part of a functional human protein ‘would represent the first report of its kind and would have important evolutionary implications’. Despite the 7 years since this challenge, confirmed cases of Alu-containing sequences that encode a functional protein still remain extremely elusive.

The paucity of documented examples is a good indication that proteins are unlikely to utilize domains encoded by Alus for functional ends. The reluctance to accept this view is understandable, given the huge proportion of interspersed repeats in the human genome (around 45%; Smit, 1999): in principle, at least some of them might have been recruited for functional purposes at the protein level. The great majority of previously detected repeat-derived coding sequences comes, however, from protein-coding repeats, and particularly from DNA transposons (Smit, 1999; Boeke, 1997). LINEs are less common in coding sequences and only a few Alus had been identified prior to the analysis of Nekrutenko and Li (2001) and Li et al. (2001). Since SINEs are derived from RNA genes without protein-coding capacity, the lack of Alu-encoded proteins is consistent with the notion that new domains arise from existing sequences encoding functional proteins (for example, by exon shuffling) and that the de novo creation of coding sequences from non-coding DNA is rare (although not impossible; see Part 2, Section 3.4).

The relative frequencies for the TE classes found by Nekrutenko and Li (2001) are similar to genome-wide repeat proportions, i.e. to expectations under random sampling of sequences or random errors in predicting exons. In contrast, our findings are in good agreement with previous reports (Smit, 1999; Lander et al., 2001) and the above arguments that repeat-derived protein-coding sequences, especially those corresponding to Alus and other SINEs, should be rare. Indeed, Alus are derived from 7SL RNA, part of the signal recognition particle on ribosomes (Ullu and Tschudi, 1984), and the strong selection for such 7SL-like secondary structures that they apparently experience (Boeke, 1997) would not leave much freedom for Alu RNAs to fulfill other roles: in particular, it would be difficult to harness them to simultaneously encode functionally important proteins.
It should be emphasized that we are not questioning the presence, or even a proposed abundance (Nekrutenko and Li, 2001), of alternatively spliced transcripts containing Alus. We are questioning the notion that such transcripts will be translated to yield functional proteins, except possibly in one or two extremely rare cases. In spite of anecdotal reports of involvement of Alu-derived or -sequestered amino acid sequences in molecular recognition or binding (e.g. of the HPK1 Alu in activating AP1; Nekrutenko and Li (2001), or of a group of Alu-derived peptides in binding tau; Hoenika et al., 2002), it is not at all clear that such possibly fortuitous involvement reflects a functional role.

**In summary,** the available examples suggest that transposable elements could occasionally correspond to parts of rare, atypical proteins that arise by alternative splicing and subsequent translation, resulting in aberrant products with potentially pathological effects. In general, however, the presence of SINEs in a putative or predicted human coding sequence still appears to be a good indication that it will seldom, if ever, be translated into a functional protein *in vivo.*
The distribution of duplicated genes in the human genome

The reason for dealing with the distribution of duplicated genes in the human genome in this part of the book is that many gene duplications are very ancient (see below) and that most duplicated genes have been transposed to other locations since.

Thirty years ago, Ohno (1970) proposed that two complete genome duplications took place at the origin of vertebrates, and that this tetraploidization was followed by the diversification of the functions of duplicated genes. Ohno’s hypothesis is also called the 2R hypothesis (for two rounds of genome duplication).

More precisely, the model proposed two duplications before and after the emergence of the jawless fishes, approximately 500 and 430 Mys ago (Ohno, 1970, 1999; Lundin, 1993; Holland et al., 1994; Sidow, 1996; Katsanis et al., 1996; Spring, 1997; Bailey et al., 1997). Two rounds of complete genome duplication should lead to a 1:4 ratio between invertebrate and mammalian coding sequences. The Hox clusters fit the model well since they occur in one copy in invertebrates and in four copies in most vertebrate genomes. In humans, potentially quadruplicated regions have been analyzed in detail, on chromosomes 1, 6, 9 and 19, by Endo et al. (1997), and on chromosomes 4, 5, 8 and 10, by Pebusque et al. (1998). Phylogenetic analysis of paralogous genes has, however, challenged this view (see Wolfe, 2001, for a review).

Since many gene duplications in the human genome are ancient duplications going back to the origin of vertebrates (see above), the question may be asked about the fate of such duplicated genes at the compositional transitions from cold- to warm-blooded vertebrates, when about half of the genes increased their GC levels. The compositional distribution of the 1,111 pairs of duplicated genes investigated by Jabbari et al. (2003) is very similar to the

Figure 6.18 Distribution of GC3 values of the 1,111 human gene pairs used (complete data set). (From Jabbari et al., 2003b).
general distribution of human genes (Fig. 6.18), where GC3-poor (<60% GC3) and GC3-rich (>60% GC3) genes are present in comparable amounts.

Two possibilities might then be considered (Fig. 6.19A), namely that the compositional transitions affected either (i) one copy of each gene pair only (or, at least, preferentially); or, in addition, (ii) none or both of the two copies, with approximately the same probability. Ideally, the two possibilities could be distinguished if each copy is put in one sub-set according to its composition. Indeed, the two distributions would be different if one copy kept the ancestral GC-poor composition, whereas the second copy, which had supposedly acquired a different function, had undergone the compositional change. In contrast, in the other case the distribution just mentioned would be blurred by the copies which both changed or maintained their composition. As shown in Fig. 6.19B, a comparison of the GC3-poor copies with the GC3-rich copies indicates that, indeed, one copy preferentially underwent the compositional transition, suggesting a different structure/function of one copy relative to the other one (Jabbari et al., 2003). Expectedly, we are not quite in the ideal case presented above. This can be understood because some recently duplicated genes are present in the data set. Indeed, if only genes showing a Nei-Gojobori (1986) distance higher than 0.7, or if only genes located on different chromosomes are taken into account (another way of eliminating recent tandem duplications), the difference between the two histograms is increased, with over 45% of the copies present in the GC-rich class.

One should now consider that (i) the two copies are generally located very far from each other in the genome (in fact they are most frequently located on different chromosomes; see Fig. 6.20A); (ii) that the compositional transition specifically affected the “ancestral genome core” (the gene-dense compartment of the ancestors of present-day mammals and birds; see Part 12); and (iii) that the GC-enriched copy is characterized by shorter introns compared to the GC-poor copy (Rayko et al., 2003; see Fig. 6.20B); the average intron size shortening by a factor of 2 is probably accompanied by an increased expression. Then, one should conclude that the “major line” of events following the ancestral gene duplications (disregarding a number of phenomena, like recent duplications, because they do not blur the events concerning the majority of duplicated genes) can be described as a succession of

Figure 6.19. A. Model of two extreme situations hypothesized for duplicated genes at the transitions from cold- to warm-blooded vertebrates. 1) One copy of each pair preferentially undergoes the transition (red arrow), the other copy maintaining its original low GC-level (blue arrow). 2) In addition to situation 1, both copies may undergo the transition or maintain their original low GC level. B. Distribution of GC3 values of the 1,111 human gene pairs used (sets of GC3-poor and GC3-rich copies). (From Jabbari et al., 2003b).
Figure 6.20. **A.** Chromosomal distribution of human gene families. Most families are localized on different chromosomes (dispersed copies), a minority being on the same chromosome. **B.** Average intron size for the GC poor and GC-rich copies of human duplicated genes. All pairs (111) for which genomic sequences were available were analyzed (From Rayko et al., 2004).

![Image](https://example.com/image.png)

Figure 6.21. A scheme of a frequent pathway following ancient gene duplications (blue bars). One copy is supposed to be preferentially transposed into the ancestral genome core (pink bar) which then undergoes the major compositional transition (red bars). (From Rayko et al., 2004).
three steps which occurred in duplicated genes during the evolution of the vertebrate genome (see Fig. 6.21).

(i) Most of the original gene duplications of interest here seem to have occurred in the fish genomes or earlier in evolution. This is indicated by the compositional divergence which is so strong as not to allow to recognise any correlation between the copies of a given gene (see Jabbari et al., 2003), whereas genes arising from recent duplications show very little compositional divergence, and a good correlation can be observed between orthologous genes from human and *Xenopus* (Cruveiller et al. 1999), or human and zebrafish (Belle et al., 2002).

(ii) While in the original duplication events, genes were tandemly arranged, as suggested by many examples of recent duplications, the preferential transposition of one copy is indicated by the fact that the majority of duplicated genes are found on different human chromosomes. It is interesting to notice that segmental duplications are enriched within pericentromeric and subtelomeric regions in the human genome (Amann et al., 1996; Trask et al., 1998, Eichler, 1999, Horvath et al., 2000); this bias has been quantitatively tested recently in the working draft of the human reference sequence (Bailey et al., 2001), and enrichment levels were 4.7- to 11.8-fold; this bias appears to be more pronounced for interchromosomal than for intrachromosomal duplications (for a review see Samonte and Eichler, 2002), and for pericentromeric compared to subtelomeric regions.

(iii) In the majority of cases, only one copy underwent the compositional change. Indeed, if a majority of both genes from gene pairs had either undergone the compositional transition, or maintained the original composition, there would be little compositional difference between the two genes of each pair, and Fig. 6.19B would not show mirror images. On the other hand, we know that compositional genome changes took place in the “ancestral genome core” of the ancestors of warmblooded vertebrates. The compositional change of one copy occurred, therefore, in the copy located in the “ancestral genome core”, whereas the other copy, maintaining the original low GC level remained in the “genome desert”, which did not undergo any compositional change.

To sum up, three steps could be identified in the major line of evolution of duplicated genes: (i) the original tandem duplications; (ii) the preferential transposition of one copy of the duplicated genes into the “ancestral genome core”; (iii) the compositional change of this copy.

These steps may suggest (i) that duplications occurred most frequently in the “ancestral genome desert”, in agreement with the preferential duplications in GC-poor pericentromeric regions (see Saccone et al., 2002); (ii) that the duplicated copy acquired, in all likelihood, a new function and/or regulation through subfunctionalization (Force et al., 1999; Avaron et al., 2003) or function partitioning (Wagner, 2002) or other mechanisms; and (iii) that transposition of the duplicated copy into the open chromatin of the “ancestral genome core” was generally preferred, as in the case of retroviral integrations (see Tsyba et al., 2004). Interestingly, these latter events led to a further gene enrichment in the “ancestral genome core”. Moreover, the expected increased hydrophobicity of the proteins encoded by genes located in the genome core were accompanied by a shortening of introns and by a preferential formation of CpG islands (Rayko et al., 2004).
Part 7
The organization of chromosomes in vertebrates
Isoches and chromosomal bands

Over thirty years ago, fluorescence staining with quinacrine was shown to produce bands on human metaphase chromosomes (Caspersson et al., 1971). This breakthrough, allowed, for the first time, the identification of individual chromosomes through their specific banding pattern and was followed by the development of other banding methods. The most widespread of these relied on treating metaphase chromosomes with dyes after proteolytic digestion or denaturation and produced G bands (Giemsa positive or Giemsa dark bands, which are equivalent to Q bands or Quinacrine bands) and R bands (Reverse bands, which are equivalent to Giemsa negative or Giemsa light bands). Since quinacrine was known to fluoresce with AT-rich DNA (Weisblum and de Haseth, 1972), from the beginning there was an indication that chromosomal bands had something to do with base composition of DNA.

Ten years later, we proposed as a working hypothesis that the G and R bands of the chromosomes from higher vertebrates might be related to isochores (Cuny et al., 1981). The reasons for such a suggestion were the following: (i) as just mentioned, there were indications (summarized by Comings, 1978) that G bands corresponded to AT-rich, late-replicating DNA, and R bands to GC-rich, early replicating DNA; likewise, light (GC-poor) and heavy (GC-rich) isochores were interspersed on chromosomes forming a compositional mosaic; (ii) G banding was very evident in warm-blooded, but not in cold-blooded vertebrates (Bailly et al., 1973; Stock and Mengden, 1975; Schmid, 1978), a feature paralleled by the high and low compositional heterogeneity of the genomes of warm- and cold-blooded vertebrates, respectively; (iii) G banding patterns appeared to be conserved in birds (Stock and Mengden, 1975) and in different orders of mammals (Dutrillaux et al., 1975; Wurster-Hill and Gray, 1979), as are the relative amounts and GC levels of major DNA components in those species; (iv) the amount of DNA per chromosome band, at the highest resolution achieved by Yunis (1976), was compatible with isochore size; (v) gene amplification leads to the formation of homogeneous staining regions in chromosomes (Schimke, 1982), the result expected if the amplified genome segments were smaller in size than isochores, which was the case.

The general suggestion of Cuny et al. (1981) seemed to be confirmed by a number of later results (Bernardi, 1989): (i) additional evidence became available for the poor banding (Abe et al., 1988; Medrano et al., 1988; Schmid and Guttenbach, 1988; Yonenaga-Yassuda et al., 1988) and for the limited compositional heterogeneity of the genomes from cold-blooded vertebrates (Bernardi and Bernardi, 1990a), as well as, for the correlation between compositional heterogeneity of DNA and intensity of chromosomal bands in fishes (Medrano et al., 1988; see Part 4, Section 1.3); (ii) G bands replicate late, R bands replicate early (Comings, 1978), as do genes previously investigated in replication timing (Furst et al., 1981; Goldman et al., 1984) and localized in GC-poor and GC-rich isochores, respectively (Bernardi et al., 1985b); (iii) genes are preferentially located in R bands.
(Korenberg and Engels, 1978; Goldman et al., 1984), as well as in GC-rich isochores (see Part 5); (iv) genes located in G and R bands are GC-poor and GC-rich, respectively (Aota and Ikemura, 1986; Ikemura and Aota, 1988), as are genes located in GC-poor and GC-rich isochores (Bernardi et al., 1985b); (v) G bands can be produced by Hae III degradation of metaphase chromosomes (Lima de Faria et al., 1980), as expected from the fact that the GGCC sites split by Hae III are much more frequent in GC-rich compared to GC-poor isochores; in contrast, R bands can be produced by pancreatic DNase degradation of chromosomes protected by GC-specific binding of chromomycin A3 (Schweizer, 1977).

I stressed, however, the fact that the general correlation between isochores and chromosomal bands could only be considered an approximation of the actual situation (Bernardi, 1989), because GC-rich (H1, H2, H3) and GC-poor (L1, L2) DNA components in the human genome are in a 1:2 ratio, whereas R and G bands are in an approximate 1:1 ratio. This discrepancy may also be due, in part, to the fact that DNA density is lower in R bands compared to G bands, an explanation in agreement with the greater DNase accessibility of some chromosomal regions (Garel and Axel, 1976; Weintraub and Groudine, 1976; Kerem et al., 1984) where gene density is higher. An alternative explanation, not exclusive of the former, is that “standard” R bands (i.e., R bands at the resolution of 400 bands), contain more GC-poor isochores (corresponding to the “thin” G bands only revealed by high resolution banding; Yunis, 1976; 1981; Viegas-Pequignot and Dutrillaux, 1978; Sawyer and Hozier, 1986) than “standard” G bands contain GC-rich isochores (corresponding to “thin” R bands). Both explanations appear to be supported by data from Ikemura and Aota (1988), by the compositional mapping of chromosome 21, and by chromosomal compositional mapping (see below).

G and R bands not only differ in their overall isochore make-up, but also in their internal isochore structure as indicated by the following points. G bands are remarkably homogeneous in DNA composition, because they are made up of GC-poor isochores that differ very little from each other in composition. In contrast, R bands are heterogeneous, since the corresponding GC-rich isochores encompass a wide GC-range. This leads to both inter- and intra-band heterogeneity, as shown by the compositional mapping of chromosome 21 (see below). In fact, the interspersion of different GC-rich isochores within individual R bands was already indicated by the finding that genes located in R bands from many chromosomes (Aota and Ikemura, 1986; Ikemura and Aota, 1988) are present in H3 isochores. Since the latter only represent 4–5% of the genome, they are far from accounting for the totality of DNA of R bands (which represent almost 50% of the DNA), and other DNA components must be present. A corollary of this conclusion was that, since gene concentration is highest in the isochores of the H3 family (Bernardi et al., 1985b; Mou chiroud et al., 1991; Zoubak et al., 1996), regions of high and low gene concentration should exist not only in R and G bands, respectively, but also within R bands.
Chapter 2

Compositional mapping

2.1. Compositional mapping based on physical maps

The problems discussed in the previous chapter obviously needed an experimental approach to be solved. This approach was compositional mapping (Bernardi, 1989), which can be described as follows. Wherever long-range physical maps are available, compositional maps can be constructed by assessing GC levels around landmarks (localized on the physical maps) that can be probed. This simply requires the hybridization of the probes on DNA fractionated according to base composition. If DNA preparations of about 100 kb in size are used, compositional mapping will provide the equivalent of a high-resolution banding, without the uncertainties of cytogenetics. Basically, this approach involved three steps: (i) DNA fractionation (see Fig. 7.1A); (ii) hybridization of single-copy sequence probes on the fractions (see Fig. 7.1B); and (iii) the construction of the map (see Fig. 7.1C).

This experimental approach was originally tried for a set of 50 single-copy sequence probes localized on the long arm of chromosome 21 and provided a direct demonstration for the compositional homogeneity of G bands and for the compositional heterogeneity of R bands, the highest GC levels (corresponding to the H3 DNA component) being in the telomere-proximal regions (Gardiner et al., 1990; see Fig. 7.1C). Incidentally, the latter observation agrees with cytogenetical evidence that telomeres almost always correspond to R bands and that the terminal regions of 20 of them (including that of the long arm of chromosome 21) are the regions (the T bands) of human chromosomes that are most resistant to heat denaturation (Dutrillaux, 1973). This suggested that H3 isochores may correspond to these denaturation-resistant telomeric regions and to some similar intercalary regions located on several chromosomes (for example, 11, 19 and 22).

The approach used for the compositional mapping of the long arm of chromosome 21 was extended to (i) the mapping of telomeric probes corresponding to T bands and to R bands, respectively, on compositional DNA fractions (De Sario et al., 1991); in this case, it could be shown that the T band probes hybridized on H3 isochores, the R band probes on H2 and H1 isochores and the telomeric repeat common to all chromosomes on H1, H2 and H3 isochores, as expected; moreover, genes localized in G, R and T bands showed increasing GC levels; (ii) the mapping of the distrophin gene (Bettecken et al., 1992); in this very large (3 Mb) gene only very small (1–2% GC) compositional differences were found to occur; and (iii) the mapping of the Xq26-qter region (Pilia et al., 1993); in this case, the difference between G and R bands first seen on chromosome 21 was also found; moreover, the telomeric regions of the Xq 28band showed very high GC levels compared not only to G bands Xq27.1 and Xq27.3 and to R bands Xq26 and Xq27.1, but also to the centromeric part of Xq28.

This result prompted a more detailed study of band Xq28 (De Sario et al., 1996), which was done by determining (De Sario et al., 1995) the buoyant densities (and GC levels) of
Figure 7.1. A Analytical CsCl profiles of human DNA fractions (P stands for Pellet), as obtained after preparative ultracentrifugation in Cs₂SO₄/BAMD at a ligand/nucleotide molar ratio Rₚ = 0.14. Relative amounts and modal buoyant densities are indicated. B. Hybridization of probes 267, H8 and CRY A on the DNA fractions from the left panel. C. Compositional map of the long arm of human chromosome 21 obtained when hybridisation data are superimposed on the physical map (Gardner et al., 1988, 1990). Long horizontal lines indicate positions of the breakpoints associated with the rearranged chromosomes listed at the right of the figure. Some regions are not labelled for lack of space. (From Gardiner et al., 1990).
Figure 7.2 Compositional map of human chromosome band Xq28. Yeast artificial chromosomes (YACs), represented as boxes with a height corresponding to 1% GC, are positioned along the physical map on the horizontal scale, and according to their buoyant density and GC level (right and left ordinates respectively) on the vertical scale. Asterisks represent GC levels of isochores calculated from GC3 values of mapped genes (also indicated by asterisks on gene names on the top part of the figure with the exception, for lack of space, of the EDMD gene). Crosses represent GC levels of isochores as evaluated by hybridization of single copy probes on compositional DNA fraction. The top part of the figure shows a number of mapped genes and loci, as well as the location of CpG islands (thin vertical lines); the region between R/GCP and DXS439 comprises seven CpG islands (unresolved in the Figure). The bottom part of the figure shows YAC contigs A-E, as well as the proximal, middle, and distal regions described in the text. (From De Sario et al., 1996).

YACs (yeast artificial chromosomes) covering this band (see Fig. 1.9); this allowed the first construction of an essentially complete compositional map of a chromosomal band, corresponding in this case to 8 Mb of DNA (Fig. 7.2). Three regions were observed: (i) a proximal 3.5 Mb region formed by GC-poor L and GC-rich H1 isochores; (ii) a middle 2.2 Mb region essentially formed by a GC-rich H2 isochore and a very GC-rich H3 isochore separated by a GC-poor L isochore; interestingly, the size of the H3 isochore is only 0.2 Mb. Gene and CpG island concentrations increased with the GC levels of the isochores, as expected. Xq28 exemplified a subset of R bands, the H3* bands which are different from the two other subsets, the H3+ bands (which are characterized by the predominance of H2 and H3 isochores and by their resistance to heat-denaturation) and the H3- bands (which do not contain H2 and H3 isochores and correspond to the majority of R bands).

A region of human chromosome 21, the 13–14 Mb cen-q21 region, was mapped using the same approach. In this case a compositional map of the centromere and of the subcentromeric region of the long arm (De Sario et al., 1997) was established by determining the GC levels of 11 YACs covering this region which extends from the α-satellite sequences of the C(centromeric) band q11.1, through R band q11.2, to the proximal part of G band q21 (Fig. 7.3). The entire region is made up of GC-poor, or L, isochores with only one GC-rich H1 isochore, at least 2 Mb in size, located in band q21. The almost identical GC levels of the centromeric α-satellite repeats (38.5%), of R band q11.2 (39%), and of the proximal part of G band q21 (38–40%) provided a direct demonstration that base composition cannot be the only cause of the cytogenetic differences between C bands, G bands and the majority of R bands, namely the H3- R bands (which do not contain the GC-richest H3 isochores). The results obtained also showed that isochores may be as long as 6 Mb, at least in the GC-poor regions of the genome, and supported previous observations made on
Figure 7.3. Compositional map of the human chromosomal region 21cen-q21. Yeast Artificial Chromosomes (YACs) are represented as boxes (whose height provides the error bar) and are positioned on the physical map according to the STS order (Chumakov et al., 1992) and on the vertical axis according to the measured buoyant densities and the calculated GC levels (right- and left-hand ordinates, respectively). The physical distances were calculated according to the integrated map of Lawrence et al. (1993). Physical distances between adjacent contigs and STSs are not at scale. YAC length are proportional to the measured size. Gaps are represented by breaks in the horizontal line at the top of the figure. The ideogram of chromosome bands in the analyzed region is shown on the top of the figure. YAC 831B6 showed a buoyant density corresponding to 44% GC (broken line box), but its actual GC level is close to 39% GC (solid line box, see text). The overlap region of YACs 759D3 and 881D2 is shaded. Asterisks represent GC levels of isochores as calculated by hybridization of single-copy probes on compositional DNA fractions (Gardiner et al., 1990). The corresponding STSs are also marked with an asterisks. The broken line in YAC 770B3 indicate the absence of STS C79. Arrows represent NotI sites as reported in the literature (Ichikawa et al., 1992; Hattori et al., 1993; Saito et al., 1991). (From De Sario et al., 1997).

Xq28 (De Sario et al., 1996) suggesting that YACs from isochore borders are unstable and/or difficult to clone. Genes and CpG islands are very rare in the GC-poor region investigated, as expected from the fact that their concentration is proportional to the GC levels of the isochores in which they are contained.

2.2. Chromosomal compositional mapping at a 400-band resolution

Compositional mapping can also be done at the chromosomal level by in situ hybridization of compositional DNA fractions corresponding to different isochore families. In the first investigation of this kind (Saccone et al., 1992), it was shown that the hybridization of a DNA fraction derived from H3 isochores produced the highest concentration of signals on two largely coincident subsets of R bands (Figs. 7.4, 7.5): (i) the T bands (Dutrillaux, 1973), which are the most heat-denaturation-resistant R bands; and (ii) the chromomycin
A3-positive, DAPI-negative bands (Ambros and Sumner, 1987), which are the GC-richest bands of human chromosomes (DAPI is 4,6 diamino-2-phenylindole). *In situ* hybridization of H3 isochore DNA established that T bands comprise GC-rich, gene-rich, single-copy DNA. Indeed the contribution of repetitive DNAs was suppressed by competition with excess unlabeled total human DNA and Alu sequences. The latter are the most abundant type of short interspersed elements, which have their highest density in the GC-richest isochores H2 and H3 (see Part 6). This also ruled out the possibility that T bands corresponded to GC-rich satellite DNAs, which was a distinct possibility until then.

In subsequent investigations (Saccone et al., 1993), the chromosomal locations of the other isochore families L1 + L2, H1 and H2 were studied by using an improved competition protocol with excess Cot 1 DNA, which eliminated the need for a statistical approach. This established (i) that T bands contain not only H3 isochores, but also H2 and some H1 isochores; (ii) that R' bands (i.e., R bands exclusive of T or H3+ bands) are formed, on the average and to almost equal extents, by H1 and L isochores, H2 and H3 isochores being only rarely present; and (iii) that G bands essentially consist of L isochores, H1 isochores being present at low levels.

In further work (Saccone et al., 1996), it was shown that (i) the number of bands

Figure 7.4. The human DNA fraction corresponding to isochore family H3 was hybridized in situ to human male metaphase chromosome spreads. a and c Detection of hybridized biotinylated H3 fraction DNA by fluorescent isothiocyanate-conjugated avidin; chromosomes are counterstained with propidium iodide. b and d G banding of metaphases shown in a and c, respectively. In a, hybridization was carried out with the H3 DNA probe at a final concentration of 20 ng/µl in the presence of a 1000-fold excess of sonicated unlabeled total human DNA to suppress hybridization of repeated sequences present in the probe. In c, a 50-fold excess of a pBR322 plasmid containing the Blur 8 sequence was added to the 1000-fold excess of competitor DNA; the results were the same in both cases. Chromosomes were subsequently G-banded with Wright's stain. (From Saccone et al., 1992).
characterized by strong H3 hybridization signals, which were called T or H3+, is 28; (ii) 31 additional R bands, called T' or H3* bands, also contain H3 isochores, yet at a lower concentration than H3+ bands; (iii) H2 and H3 isochores appear to have, as a rule, the same band distribution; and (iv) the remaining R bands (about 140 out of 200, at a resolution of 400 bands), called H3- bands, do not contain any detectable H3 isochores (Fig. 7.6).

The existence of three distinct sets of R bands was further supported by (i) the different compositional features of genes located in them; (ii) the very low gene density of chromosomes 13 and 18, in which all R bands are H3+ bands; (iii) the compositional map of a H3* band, Xq28, and of a H3- band, 21q11.2 (see the preceding section); (iv) the predominance of GC-rich and GC-poor long (>50 kb) sequences in H3+/H3* and in H3-/G bands, respectively; and (v) the large degree of coincidence of H3+ and H3* bands with CpG island-positive bands; it should be stressed that hybridization of CpG island-enriched DNA (Craig and Bickmore, 1994) leads to poor patterns, in which bands can not be identified, except by comparison with the bands obtained by isochore hybridization (as done by Saccone et al., 1996).
2.3. Chromosomal compositional mapping at a 850-band resolution

The results described so far were obtained at the standard resolution of 400 bands. Using prometaphase chromosomes, a higher resolution of 850 bands can be achieved. In this case, the distribution of hybridization signals on chromosomes (Fig. 7.7) showed that H3 isochores are only located on a small number of $R_{850}$ bands (Saccone et al., 1999; Fig. 7.8; $R_{400}$, $G_{400}$, $R_{850}$, $G_{850}$ indicate the R and G bands at resolutions of 400 and 850 bands, respectively). In fact, H3 hybridization signals covered almost all, many, and only a few of the $R_{850}$ bands derived from $R_{400}$ $H3^+$ bands, $H3^*$ bands, and $H3^-$ bands, respectively (Fig. 7.9).
Indeed, 23 out of the 28 R400 H3\(^+\) bands only yielded R850 bands containing H3 isochores, whereas only some of the R850 bands originating from the other five R400 H3\(^+\) bands showed H3 hybridization signals. For example (see Fig. 7.9), the R400 H3\(^+\) band 11q13 is resolved into three R850 H3\(^+\) bands (q13.1, q13.3, and q13.5), and two G850 bands (q13.2, and q13.4), whereas the H3\(^+\) band 11p15 was one of the five exceptions, with only one (11p15.5) of the three derived R850 bands showing hybridization signals. Globally, the coverage of the R850 bands derived from the 28 R400 H3\(^+\) bands was 91\% (this calculation was based on band sizes as assessed by Francke, 1994). In no case G850 bands derived from the R400 H3\(^+\) bands showed hybridization signals.

In contrast, only some of the R850 bands derived from 23 out of 31 R400 H3\(^*\) bands showed hybridization signals. For example, the H3\(^*\) band 11q23 (Fig. 7.9) yielded only one R850 H3\(^*\) band (11q23.3; in fact, only the distal part of it was H3\(^+\)), whereas the other R850 band (11q23.1) was H3\(^-\). The remaining eight H3\(^*\) bands showed the features observed in the vast majority of H3\(^+\) bands, in that all the derived R850 bands were H3\(^+\). Globally, 66\% (in terms of band sizes; see above) of the R850 bands derived from the R400 H3\(^*\) bands contained H3 isochores.
Figure 7.8. Ideogram of human chromosomes at a 850-band resolution (Francke, 1994) showing the H3' bands as red bands. (From Saccone et al., 1999).
As far as the R_{400} H3⁻ bands are concerned, the higher resolution allowed the identification of 20 bands containing H3 isochores that had not been detected at the lower resolution (Saccone et al., 1996). The majority of these bands were located close to other H3⁺ or H3* bands (see bands 5q33, 1, 6p21.1, and 12q24.13) and were very thin (see bands 1p13.3, and 7p13). Only some of the R_{850} bands derived from these 20 R_{400} H3⁻ bands (see Fig. 7.8) exhibited hybridization signals (see band 11p11.2 in Fig. 7.9). Moreover, in a number of cases, the signal was thinner than the corresponding R_{850} bands (see Fig. 7.7), indicating that only part of the R_{850} band contained H3 isochores. Globally, only 9% (in terms of size; see above) of the R_{850} bands derived from the H3⁻ bands contained H3 isochores. If only the R_{850} bands derived from the 20 H3⁻ bands were taken into consideration, the coverage was 47%. Incidentally, previous work (Saccone et al., 1996) had shown that H2 and H3 isochores co-localize on metaphase chromosomes, with only four exceptions (the telomeric bands 3q29, 6q27, 13q34, and 20p13) which were H2⁺ and H3⁻. Now, these bands were shown to be H3⁺, indicating also in these cases a co-localization of H3 and H2 isochores.

On the basis of the above results and of the estimated band sizes (Francke, 1994), it can be calculated that about 17% of all bands at a 850-band resolution contain H3 isochores.
and that about 9%, 6%, and 2% of them derived from $R_{400}$ H$^+$, H$^*$ and H$^-$ bands, respectively.

Finally, G bands did not reveal the presence of H3 isochores, the only exceptions being two $G_{400}$ bands, 1p36.2 and 19q13.4, which yielded two $R_{850}$ H$^+$ bands, 1p36.22, and 19q13.42, respectively (Fig. 7.8).

**In summary,** the results just presented lead to several conclusions. (i) Since the colocalization of H2 and H3 isochores (which represent together 12% of the human genome) in $R_{850}$ H$^+$ bands appears to be the rule, the fraction of these isochores in those bands (which represent 17% of the total genome) corresponds to the majority, 70%, of the DNA contained in them; this value should, however, be higher since DNA is less compact in these bands (see Chapter 3). (ii) In some cases, however, the coverage of $R_{850}$ H$^+$ bands by hybridization signals is overestimated. For example, signals suggest that almost 50% of band Xq28 is H$^+$, whereas compositional mapping showed that only about 25% is formed by H2/H3 isochores (De Sario et al., 1996). (iii) In a number of $R_{850}$ H$^+$ bands (see Fig. 7.8), H3 hybridization coverage was limited to a fraction of the band (see for example the case of band Xq28); this indicates that the present results provide information concerning a resolution higher than 850 bands; thus, they may correspond, in many cases, to the practical highest resolution that can be reproducibly attained, namely 1,250 bands (Drouin et al., 1994; Scherthan et al., 1994). (iv) 83% of the bands shown in Fig. 7.8, namely the $R_{850}$ H$^+$ and the $G_{850}$ bands, exhibit low or very low gene concentrations; since genome size is remarkably constant in mammals and since such regions are conserved in syntenic regions of chromosomes from mammalian orders that diverged about 100 million years ago (Saccone et al., 1997; Scherthan et al., 1994; Raudsepp et al., 1996; Morescalchi et al., 1997; Chowdhary et al., 1998), the gene-poor majority of the genome should have some functional role (see Part 12). (v) Finally, the gene-richest regions correspond to gaps in the first physical map of the human genome (Chumakov et al., 1995), which therefore missed the most interesting regions of the human karyotype (as already stressed by Saccone et al., 1996). The difficulty experienced in cloning these regions into YACs and/or in avoiding high levels of chimerism and deletion is most probably related to their high recombination level, a property which apparently is conserved when these regions are cloned in yeast.

The analysis of the localization of GC-poor DNA on metaphase and prometaphase chromosomes will be illustrated in some detail in order to provide an example of the procedure followed. Fig. 7.10 displays the CsCl profiles, the relative DNA level, and the proportion of the isochores families in the DNA fractions (Saccone et al., 1996) used as probes for the *in situ* hybridizations of GC-poor DNA. L1 isochores are only present in the pellet DNA; L2 isochores are distributed in the pellet and in fractions 1–3; fractions 4 and 5 contain almost exclusively DNA from H1 isochores. In the case of fraction 5 (as well as in the following fractions shown in Fig. 1 of Saccone et al., 1996), a light satellite peak corresponding to DNA from centromeric heterochromatin is present. The pellet DNA, essentially formed by L1 isochores, hybridized on a subset of G bands (Fig. 7.11), which were called L1 bands, and corresponded to the G1 and G2 bands, as we will name the two most intensely staining subsets of Francke’s R bands (1994). In contrast, the pellet DNA is almost absent in the large majority of the H$^{+}_{400}$ and H$^{*}_{400}$ bands. This is especially evident in chromosomes 15, 17, 19, and 22 (see Fig. 7.11).
Figure 7.10. CsCl profiles of the human DNA fractions containing the L1, L2 and H1 isochore families (modified from Saccone et al., 1996). Fraction numbers (P stands for pellet), relative DNA content, modal buoyant density and relative amounts of isochore families are indicated. (From Federico et al., 2000)

Figure 7.11. L1 isochore banding pattern. The left and right member of each chromosome pair shows the GTG (G bands by trypsin using Giemsa) bands and the L1 isochore hybridization, respectively. Biotinylated L1 isochores present in the pellet DNA shown in Figure 7.10 were detected with avidin-FITC. Chromosomes were stained with propidium iodide. (From Federico et al., 2000).
Interestingly, the pellet DNA represents about 40% of the total genome, most of which belongs to the L1 isochores, whereas the bands on which it is localized correspond to about 26% of all bands (as assessed from the ideogram of Francke, 1994). This indicates that the level of DNA compaction in L1 bands is higher than the average (see below). The high resolution hybridization obtained with pellet DNA gave rise to signals characterized by

Figure 7.12. Identification of the GC-poorest and the GC-richest chromosomal bands. Human karyotype at a resolution of 850 bands showing the chromosomal bands containing the GC-poorest (L1\(^{\pm}\), blue bars) and the GC-richest isochores (H3\(^{\pm}\), red bars). The H3\(^{\pm}\) R bands are in white. The grey scale of the G bands is according to Francke (1994). (Modified from Federico et al., 2000).
different features. The majority of L1 bands was present in doublets separated by thin non-hybridizing internal bands (see Fig. 7.12). The G1 and G2 bands contain almost the totality of the GC-poorest isochores present in the pellet. More precisely, about 85% of the G1 bands and about 40% of the G2 bands showed strong signals, the remaining G2 and G1 bands showing very weak or no signals (see Fig. 7.12). Only few (about 4%) G3 bands and no G4 bands seem to contain L1 isochores. These are clear indications that G1+G2 and G3+G4 bands form two compositionally different subsets of G850 bands. Finally, some G1 bands do not contain pellet DNA. These bands are 1q41, 2q36.3, 3p24.3, 4q34.3, 5q34, 8p22, 10q23.1, 10q25.1, 13q33.1, 13q33.3, and 14q12. It is possible that L1 isochores are also present in these G1 bands, but in very small amounts. These bands appear to be a small, distinct subset of G1 bands that do not replicate at the end of the S phase of the cell cycle (see Chapter 6).

Table 7.1 summarizes the classification of chromosomal bands presented in Sections 2.2 and 2.3.

### TABLE 7.1
Summarizes the classification of chromosomal bands presented in Sections 2.2 and 2.3.

<table>
<thead>
<tr>
<th>Bands</th>
<th>Isochores</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>At 400-band resolution</td>
<td>R</td>
<td>T bands</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R' bands</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G bands</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>H3* or T bands</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H3* or T' bands</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H3</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>G bands</td>
</tr>
<tr>
<td>At 850-band resolution</td>
<td>R</td>
<td>H3+ bands</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H3 bands</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>L1 bands</td>
</tr>
<tr>
<td></td>
<td></td>
<td>L1* bands</td>
</tr>
</tbody>
</table>

Saccone et al., 1993
Saccone et al., 1996
Saccone et al., 1999
Federico et al., 2000
Genes, isochores and bands in human chromosomes 21 and 22

Chromosomes 21 and 22 are very different from each other. Together, they exhibit the whole spectrum of chromosomal bands. While chromosome 21 is made of several compositional regions representing all isochore families, chromosome 22 is essentially formed by H2 and H3 isochores, comprises H1 isochores, but contains very few L2 isochores and no

![Figure 7.13. A. Correlation between chromosomal bands, isochores, and gene concentration of human chromosomes 21 and 22. Bottom to top: 1) Bands, ideogram at a resolution of 850 bands showing the four classes of G bands characterized by different staining intensity (Francke, 1994; called here G1 to G4 from black to pale grey, see Fig. 7.14); and the two classes of R bands (Saccone et al., 1999; H3⁺, red; H3⁻, white); the two chromosomes are represented according to their cytogenetic size in the karyotype of Francke (1994). 2) GC, the GC profiles obtained using a window size of 100 kb; 37%, 41%, 46%, and 53% GC were taken as the upper values for the L1, L2, H1 and H2 isochore families, respectively (Zoubak et al., 1996); the grey boxes indicate DNA sequences not yet available. 3) Genes/Mb, gene density calculated as number of genes per Mb; the blue bar plot concerns chromosomal bands, the red plot 1-Mb segments. B. Compositional features of chromosomal bands. Bottom to top: 1) Bands, band ideograms as in Fig. 7.13. G1-G4, H3⁺, H3⁻ bands are indicated. 2) GC% average GC level of each chromosomal band (horizontal blue lines), and GC levels observed at band borders (red and blue arrows indicate the GC level on the R and G band side, respectively; vertical red lines indicate the GC difference over 300 kb regions around band borders). Note that all G bands showed lower GC levels than the adjacent R bands, and all R bands showed higher GC levels than the adjacent G bands. These differences were enhanced at band border regions (see above). (From Saccone et al., 2001)
L1 isochores (Fig. 7.13). Compositional fluctuations are remarkably high in the GC-richest regions compared to the GC-poorest regions. It is possible that the increase in compositional fluctuations may be due to the simultaneous requirement for AT-rich regions to link DNA to the chromosome scaffold and for the very high GC levels of coding sequences and associated CpG islands in H2 and H3 isochores.

Each band type is characterized by a defined isochore content: H3+ and L1+ bands are almost only composed of H2/H3 isochores and of L1/L2 isochores, respectively. The compositionally intermediate L1− and H3− bands are more heterogeneous, containing different proportions of L2, H1 and H2 isochores, with a relatively large amount of H1 isochores.

Matching of compositional regions with chromosomal bands (at a 850-band resolution; see Figs. 7.13A and B) was done using three criteria: the average GC level, the discontinuities in GC level at each band border (see below), and the location of genetic markers assigned to chromosomal bands.

Our identified band borders were in good agreement with the high resolution mapping of chromosome 22 markers of Kirsch et al. (2000), only two exceptions out of 19 markers being observed on contiguous bands but very near the band border. The average GC level of the bands increased from L1+ (G1+G2) to L1− (G3+G4) among the G bands, and from H3– to H3+ bands among the R bands. Interestingly, while the GC levels of H3+ bands (47–52% GC) are always higher than those of L1+ bands (35–36% GC), in the case of “intermediate” bands, G bands can be endowed with higher GC levels than those of (non-contiguous) R bands. Indeed, L1− G bands of chromosome 22 are GC-richer than most H3− R bands of chromosome 21. This situation was also found in bands from the same chromosome (compare 21q22.12 and 21q22.2 with 21q11.2 and 21q21.2). The average GC level of a G band is, however, always lower than those of the adjacent R bands, and conversely the average GC level of an R band is always higher than those of the contiguous G bands (Fig. 7.14). Moreover, the GC levels at each band border (over regions of about 300 kb) are always higher on the R side compared to the G side. Whether the situation just described for chromosomes 21 and 22 is general, remains to be established. This requires, however, a preliminary work to define satisfactory correlations between chromosomal bands and DNA sequences.

Fig. 7.13 shows bar plots of gene densities for chromosomal bands, as well as for 1-Mb regions. These data show gene density ratios as high as 20 between H3+ and L1+ bands, and as high as 30 between some 1-Mb regions. The data were used to illustrate the correlation between gene density and GC level of chromosomal bands. This is best described by two regression lines with different slopes that intersect at 42–43% GC (Fig. 7.15), in agreement with previous results obtained on DNA molecules in the 100-kb size range using an independent approach (Zoubak et al., 1996; see Fig. 5.1).

Chromosomes 21 and 22 are quite different not only in their composition, but also in their “cytogenetic” size. While the DNA amount of the long arms of both chromosomes is comparable (the difference being less than 3%, as defined by Hattori et al., 2000, and Dunham et al., 1999), the cytogenetic size of chromosome 22 is about 40% greater than that of chromosome 21, as defined at a resolution of 850 bands (Francke, 1994), indicating that DNA is more compact in chromosome 21. While it is possible that the size difference between these two chromosomes was overestimated, similar results were obtained when
In chromosomes 21 and 22, G bands are always flanked by GC-richer bands, whereas R bands are always flanked by GC-poorer G bands. As shown in the figure, G bands may, however, be even GC-richer than other (non-contiguous) R bands.

This analysis was extended to the band level. Indeed, the L1+ band 21q21.1 and the H3+ band 22q13.31 show a similar cytogenetic size and a very different DNA sequence size (about 9 Mb and 6 Mb, respectively), indicating a 50% higher compaction in the L1+ bands compared to H3+ bands. Likewise, we could estimate about 20% more and 14% less DNA than the average in the L1+ band 21q21.1 and H3+ band 21q22.3, respectively. Even if the quantitative aspects of this different compaction are subject to refinement, it is most unlikely that the qualitative conclusion about differential DNA packing is incorrect. These results provide, therefore, independent additional support, that the chromatin of GC-rich.
and gene-rich chromosomal regions has a more "open" structure compared to that of GC-poor and gene-poor regions at prometaphase.

Finally, Table 7.2 presents the classification, the relative amounts and the average gene densities of bands from chromosomes 21 and 22.

### TABLE 7.2
Classification, relative amount, and gene densities of bands from chromosomes 21 and 22.

<table>
<thead>
<tr>
<th>Bands</th>
<th>% by staining properties</th>
<th>by isochore content</th>
<th>gene density</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td>13.7%</td>
<td>L1+</td>
<td>26.3%</td>
</tr>
<tr>
<td>G2</td>
<td>12.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G3</td>
<td>13.1%</td>
<td>L1 -</td>
<td>20.7%</td>
</tr>
<tr>
<td>G4</td>
<td>7.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rev.</td>
<td>53%</td>
<td>H3 –</td>
<td>35.6%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bands</th>
<th>% by staining properties</th>
<th>by isochore content</th>
<th>gene density</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>≈15%</td>
<td>H3 +</td>
<td>17.4%</td>
</tr>
</tbody>
</table>

---

a From Saccone et al. (2001)
b Relative amounts of different band types were assessed on the basis of band sizes in the 850-band karyotype of Francke (1994).
c Here, we call G1-G4 the G bands characterized by four levels of grey (from black to pale grey) as defined by Francke (1994). T bands are the most heat denaturation-resistant R bands of Dutrillaux (1973). Relative amounts of G bands are estimated from Francke (1994), those of T bands from Dutrillaux (1973).
d As defined in Saccone et al. (1996, 1999) and Federico et al. (2000)
e This is the average number of genes per Mb found in the bands of chromosomes 21 and 22.

On a different chromosome a very informative analysis of a 17 Mb stretch covering the sequence of human Xq25 and its flanking regions was reported by Ross (2003) and is presented in Fig. 7.16. Band Xq25 was selected because it is one of the darkest G bands in the classification of Francke (1994) and because it strongly hybridized to L isochores (Federico et al., 2000). Several features are of interest: (i) the GC increases at both borders of Xq25; (ii) the inversely correlated LINE and SINE distribution; and (iii) the isochore structure of the region obtained from the IsoFinder segmentation results (Oliver et al., 2001). As expected, the gene count shows a great variation, from 0 to 14 genes per Mb.
Figure 7.16 Analysis of the sequence of human Xq25 and its flanking regions. The G-banding pattern is represented at the top. A selection of genes is shown on the cytogenetic map. These genes are linked by dotted lines to the 1-Mb sequence intervals in which they lie. The upper histogram shows the percentage of each 100-kb sequence window that is accounted for by LINE or Alu elements. The lower histogram shows the G+C content of 100-kb windows (scale in Mb). Horizontal lines show the upper limits designated for L1, L2 and H1 isochores. The two gaps between the three sequences are shown as vertical white lines. Below the G+C plot are the gene counts for each 1-Mb interval. At the base of the figure is the suggested isochore structure of the region taken from the IsoFinder segmentation results. Thin vertical lines indicate predicted boundaries between isochores of the same family. (From Ross, 2003).
CHAPTER 4

Replication timing, recombination and transcription of chromosomal bands

4.1. Replication timing of R and G bands

Chromosome replication is bimodal in all vertebrates. In the case of human chromosomes, the early investigations of Dutrillaux et al. (1976) and Biemont et al. (1978) showed that not only R and G bands differed in replication timing, R bands replicating early, G bands replicating late in the cell cycle, but also that different R bands and different G bands showed different replication timings. In fact, 18 replication timings were identified, at a 300-band resolution. As already mentioned, we identified (at a 400-band resolution) three sets of R bands: H3+, H3* and H3−, endowed with large, moderate and no detectable amount of the gene-richest H3 isochores, respectively. When the replication timings of these three sets of bands were compared with each other, it appeared that the chromosomal bands containing H3 isochores were replicating almost entirely (in the case of H3+ bands) or largely (in the case of H3* bands) at the onset of the S phase, whereas chromosomal bands not containing H3 isochore (H3− bands) were replicating later (Federico et al., 1998). The existence, at a resolution of 400 bands, of at least three distinct subsets of R band is, therefore, not only supported by their GC and gene concentration, but also by their replication timings (see Fig. 7.17).

As already mentioned, G bands comprise two different subsets of bands, one of which is predominantly composed of L1 isochores, has a higher DNA compaction relative to H3+ bands and corresponds to the darkest G bands of Francke (1994). In contrast, the other subset is composed of L2 and H1 isochores, has less extreme compositional properties and corresponds to the less dark G bands of Francke (see Fig. 7.18). When we checked the replication timing of G bands using the data of Dutrillaux et al. (1976) and Biemont et al. (1978), we found that the LI+ bands (or GI+G2; see Table 7.1) comprised the bands that replicate at the latest times, whereas the LI− bands (or G3+G4; see Table 7.1) were the earliest replicating among the G bands (Federico et al., 2000). Fig. 7.19 summarizes all the results just presented.

The best assessment of replication timing in chromosomes 11q and 21q was given by Watanabe et al. (2002; see Fig. 7.20). These authors found that replicons are heterogeneous in size (most being 50–450 kb) and that several to 10 (or more) contiguous replicons with origins that fire synchronously at a specific time comprise a megabase-size domain that can be visualized cytogenetically as a band by the replication-band method. Watanabe et al. (2002) also showed that the frequency of single nucleotide polymorphisms (SNPs) was high in the late-replicating, GC-poor regions (see Fig. 7.20) and that disease-related genes are concentrated in timing-switch regions, namely at isochore borders. This latter point is not surprising in view of a vast literature indicating that R-bands (which are characterized by frequent compositional discontinuities) and G/R borders are the predo-
Figure 7.17. Replication timing of R bands. G-banded human karyotype showing the R bands identified by Saccone et al. (1996) as H3+ (red), H3* (yellow) and H3 (white). G bands are black. The replication class of R bands of Dutrillaux et al. (1976) are indicated (by Arabic numerals) on the left (H3+ and H3* bands) and on the right (H3 bands) of each chromosome. In some cases, the same R band corresponds to two different replication timings. (From Federico et al., 1998).
Figure 7.18. Replication timing of the G bands. Human karyotype at a resolution of 400 bands per haploid genome showing the chromosomal bands containing the very GC-poor isochores present in the pellet. Blue, black, and white regions indicate the G bands containing the GC-poorest, L1, isochores, the other G bands, and the R bands, respectively. The numbers on the right and on the left of each chromosome indicate the replication timing of the corresponding regions, as described previously (Dutrillaux et al., 1976; Biémont et al., 1978). (From Federico et al., 2000).
Figure 7.19. Histograms with the replication timing of the chromosomal bands. Histograms showing the human chromosomal bands as distributed in the replication classes described by Dutrillaux et al. (1976) and Biémont et al. (1978). (From Saccone and Bernardi, 2001).

The replication timing of the chromosomal bands

minant sites of exchange processes, including spontaneous translocations, spontaneous and induced sister-chromatid exchanges, the chromosomal abnormalities seen after X-ray and chemical damage, and fragile sites (see Bernardi 1989, 1995, for reviews).

4.2. Recombination in chromosomes

According to a recent review (Nachman, 2002) “Across the genome, recombination rate varies from low values near 20 to high values close to 5 cM/Mb (centimorgans per megabase). In general, recombination appears to be low near the centromeres of metacentric but not acrocentric chromosomes. Recombination is generally elevated near the ends of chromosomes, regardless of whether or not there is a centromere in proximity” (see Fig. 7.21). The high level of recombination near the ends of chromosomes may be associated with their generally high GC levels, or, as suggested by Bernardi (1989, 1993b, 1995), with the high levels of compositional fluctuation exhibited by GC-rich isochores. Chromosomal rearrangements have two important consequences; the activation of oncogenes by strong promoters that have been put upstream of them by the rearrangement, and the possibility that some chromosome rearrangements lead, in evolutionary time, to reproductive barriers and speciation. It was speculated that the higher incidence of cancer and rate of speciation shown
Figure 7.20. Replication timing and GC% and SNP distributions on chromosome 11q. A. GC% distribution of the 11q sequence obtained from http://genome.cse.ucsc.edu/ was calculated with a 500 kb window and a 50 kb step. Undetermined bases (Ns) were omitted from the GC% calculation, but the nucleotide positions were not changed. If a total of \( N \) bases in a window exceeded a half of the window size, the data point was omitted and the neighboring data points were connected by a line. Early and late loci are denoted by red and blue ovals, respectively. Distribution of SNP frequency was calculated with a 200 kb window and a 20 kb step and is indicated by a bar diagram. A megabase-sized zone that is composed primarily of STSs belonging to one timing category is indicated by a horizontal colored line; the \( S1 \) and \( S2 \) zones are indicated by red and pink horizontal lines, respectively, and the \( S3 \) and \( S4 \) zones are indicated by light- and dark-blue lines, respectively. Chromosomal band numbers were assigned by referring to NCBI OMIM, UniSTS, and GENATLAS Sequences in the subtelomeric region were found to replicate late even though their GC% levels were rather high. B. An ideogram of the 11q band pattern reported by Saccone et al. (1999) and Bernardi (2000a). Red indicates bands that are composed mainly of the most GC-rich H3 isochores and thus correspond to T bands. Ordinary R bands are indicated in white. G bands composed primarily of AT-rich L isochores are indicated in black, and those composed of both L and H isochores are indicated in grey. (From Watanabe et al., 2002)

by warm-blooded relative to cold-blooded vertebrates correlate with the higher propensity to chromosome rearrangements of the former due to the larger number of compositional discontinuities in their chromosomes (Bernardi, 1993b).
Figure 7.21. Recombination rate variation for chromosome 10 based on data from Kong et al. (2002). The top panel for each chromosome depicts a plot of genetic versus physical position for microsatellite markers; the slope of this curve provides an estimate of recombination rate. The bottom panel gives the recombination rate estimates from Kong et al. (2002) for each marker based on a 3 Mb window. (From Nachman, 2002).

4.3. Transcription of chromosomal bands

As far as transcription is concerned, over 2.45 million SAGE (Serial Analysis of Gene Expression) transcript tags, including 160,000 tags of neuroblastomas, were investigated for 12 tissue types (Caron et al., 2001). SAGE is based on the extraction of a 10-base pair tag from a fixed position in each transcript and the sequencing of thousands of these tags and can quantitatively identify all transcripts expressed in a tissue or cell line (Velcutescu et al., 1995). Algorithms were developed to assign these tags to UniGene clusters and their chromosomal positions. The resulting Human Transcription Map generated gene expression profiles for any chromosomal region in 12 normal and pathologic tissue types (Caron
The map revealed a clustering of highly expressed genes to specific chromosomal regions. Such clusters were called RIDGES (Regions of Increased Gene Expression) by the authors, who found a correlation between gene expression and density of mapped genes for 50–60% of the RIDGES. Since gene density increases with increasing GC, one should, therefore, expect a correlation between gene expression and GC levels of RIDGES, or GC levels of isochores comprising the genes under consideration (as reported by D'Onofrio, 2002). Caron et al. (2001) were apparently not aware of our investigations, because they could have found a correlation between their RIDGES and GC-rich isochores. A surprising result reported by Caron et al. (2001) was that gene density decreases in telomeric regions, since we had shown that most telomeric regions are GC-rich and gene-rich.
5.1. Distribution of the GC-richest and GC-poorest isochores in the interphase nucleus of human and chicken.

The in situ hybridization of the human GC-richest fraction showed a general distribution of the signals in the central region of interphase nuclei, whereas the GC-poorest fraction was detected prevalently in the peripheral part (Fig. 7.22A), all the observed nuclei showing this type of signal distribution (Saccone et al., 2002). Also the hybridization of the chicken GC-richest and GC-poorest DNAs indicates that the former are localized in the interior of the nucleus, whereas the GC-poorest ones are localized in the periphery (Fig. 7.22B). The chicken nuclei also showed some large H4 isochore-hybridizing peripheral regions that could be due to the heterochromatin of the Z chromosome (macrosatellite pFN-1; see Hori et al., 1996), which was incompletely competed out by the suppression procedure (see Andreozzi et al., 2001).

5.2. Different compaction of the human GC-richest and GC-poorest chromosomal regions in interphase nuclei

To understand better, at the nuclear level, the chromatin conformation as related to the GC levels of the bands, we investigated in detail (see Fig. 7.23) the GC-richest bands 6p21 (probe A) and 9qter (probe B), the GC-intermediate bands 9pter (probe C) and 12q22–23 (probe D), and the GC-poorest band 12q21 (probe E). The GC-profiles shown in Fig. 7.23 were obtained using the draft human genome sequence (Lander et al., 2001), as redrawn by Pavliček et al. (2002a) and correlated to chromosomal bands using the approach of Saccone et al. (2001). The band DNA probes, together with the corresponding chromosome-specific DNA, were first hybridized on metaphase chromosomes to check the precise localization of the probes on the chromosomes (see Fig. 7.23) and to show that no cross-hybridization on other chromosomal sites could be detected with any of the five probes.

Each of the five chromosomal regions (Fig. 7.23, probes A to E) were then hybridized on the cell nuclei and the extensions of the hybridization signals in mitotic and interphase cells were compared. The GC-rich 6p21 and 9qter regions showed a wider spreading of the hybridization signals in the nuclei (Fig. 7.24A and B) relative to that of the chromosome territories. Moreover, a comparison of the same hybridization on metaphase (Fig. 7.23A and B) and interphase chromosomes (Fig. 7.24A and B) showed that the extensions of the two band DNAs are clearly more spread out in the nuclei. In contrast, the GC-poor 12q21 region showed a high level of compaction in the nuclei (compare Figs. 7.23E and 7.24E). In fact, the comparison of the hybridizations in the metaphase and interphase chromosomes
clearly indicated that hybridization signals of the band-specific DNA are more condensed in the nuclei compared to the mitotic cells.

The compositionally intermediate region 9pter (comprising H3 and L1 bands) showed a compaction degree comparable to that observed with the GC-poor band 12q21, the large majority of the nuclei exhibiting the hybridization signal distribution exemplified by Fig. 7.24C. Instead, the other GC-intermediate band, 12q22-23 (Fig. 7.24D), showed a signal distribution in the nuclei characterized by features intermediate between those of the GC-rich and GC-poor bands, although a number of nuclei (about 25%) were observed with more compact signals, similar to those of the GC-poor band 12q21. In any case, the two GC-intermediate bands tested never showed signals similar to those obtained for the GC-rich bands. It should be noted that the two GC-intermediate regions are located in close contiguity to bands of the GC-poorest class, suggesting that GC-intermediate bands are characterized by compaction features similar to those of the adjacent bands.

The different compaction level of the GC-rich and the GC-poor chromosomal regions is also evident by comparing their hybridizations in the nuclei. In fact, in spite of the fact that the GC-rich 6p21 and the GC-poor 12q21 bands comprise a similar DNA amount (15–20 Mb; see Fig. 7.24, probes A and E), the spreading of the signals in the nuclei is much wider.
Figure 7.23. Band DNA probes used to demonstrate the differential chromatin compaction. Top: Ideograms of human chromosomes 6, 9, and 12 at 850 band resolution, and corresponding GC profiles. The ideograms show the H3\(^+\) bands in red and the L1\(^+\) bands in blue. The GC profiles (from Pavlicek et al., 2002a) shown on the right of each chromosome are differentially coloured according to the GC level of the isochore families (L1, blue: < 37%, L2, pale blue: 37% – 41%, H1, yellow: 41% – 46%, H2, orange: 46% – 52%, H3, red: > 52%). The chromosomal bands were matched with the GC profiles according to Saccone et al. (2001). A, B, C, D, and E indicate the probes used for the \textit{in situ} hybridizations. The grey regions in the GC profiles are unsequenced regions. Bottom: Chromosomes 6, 9 and 12 showing the hybridization with the biotin-labeled probes A to E. Chromosomes were painted with digoxigenin-labeled probes. Detection was performed as in Fig. 7.22. Band type indicates the compositional features of the hybridized bands (at a resolution of 850 bands). Probes A and B detected regions mostly consisting of the GC-richest H3\(^+\) bands; probes C and D detected regions composed by GC intermediate L1\(^-\) and H3\(^-\) bands, but adjacent to GC-poorest bands; probe E detected regions mostly consisting of the GC-poorest L1\(^+\) bands. (From Saccone et al., 2002)

for the former compared to the latter (see Fig. 7.25). The results obtained with the intermediary band 12q 22–23 are also shown in Fig. 7.25 for the sake of comparison.

Concerning the nuclear location of the DNA belonging to bands endowed with different compositional features, the results are in agreement with the nuclear distribution of the GC-richest and the GC-poorest isochores (Fig. 7.24). In fact, we observed a more peripheral localization of the GC-poor bands compared to a more internal position of the GC-
rich bands (see Fig. 7.22). In the case of chromosome 9, this indicates that different regions of the same chromosome contribute to the DNA located both in the interior and the periphery of the nucleus. Moreover, at a resolution of 850 bands, probes A and B (from the two GC-richest chromosomal regions) and probe E (from one of the GC-poorest regions), detected not only the H3+ and the L1+ bands, respectively, but also the L1- and the H3- bands, respectively, that are located between them. This indicates that
Figure 7.25. Differential compaction of chromatin in human nuclei. Comparison of hybridizations obtained with the GC-richest and the GC-poorest DNA regions (partial nuclei from Fig. 7.24). The band DNA (red signals) and the corresponding chromosome (green signals) are indicated. Nuclei were DAPI stained. The very GC-rich 6p21 band showed a much larger extension compared to the very GC-poor 12q21. Furthermore, the former is more internal compared to the latter. The results obtained with the intermediary band 12q22-23 are also shown for the sake of comparison. Bar: 2 μm. (From Saccone et al., 2002).

DNA from these very small compositionally intermediate L1⁻ and H3⁻ bands also occupy the same nuclear location as the contiguous GC-richest or GC-poorest bands.

5.3. The spatial distribution of genes in interphase nuclei

In conclusion, we have shown that the GC-richest and the GC-poorest DNA of warm-blooded vertebrates (more precisely the DNA characterized by GC levels higher than 50% and lower than 38%), are located in internal and peripheral regions of the nucleus, respectively. The importance of this observation is due to the fact that a number of structural and functional features of the genomes of warm-blooded vertebrates, such as gene density, are correlated with GC levels. Our DNA probes from H3 and L1 isochores detected two subsets of chromosomal bands, namely the H3⁺ and the L1⁺ bands, which largely overlap with the T bands of Dutrillaux (1973) and with the two darkest sets of G bands of Francke (1994), respectively. On the other hand, we found that the GC intermediate band DNAs (L1⁻ and H3⁻) are characterized by a nuclear location similar to those of the L1⁺ or H3⁺ bands that are close to them on mitotic chromosomes. This accounts for the observation (Croft et al., 1999) that chromosome 18 (containing L1⁺, L1⁻ and H3⁻ bands, but no H3⁺ bands) shows a peripheral nuclear localization, whereas chromosome 19 (which contains H3⁺, L1⁻ and H3⁻ bands, but no L1⁺ bands) is located in the nuclear interior. In other
words, these exceptional situations are due to the fact that these two very small chromosomes lack the GC-richest and the GC-poorest bands, respectively.

In contrast, the large majority of chromosomes comprise both categories of bands (Federico et al., 2000; see also Fig. 7.26). It is, therefore, understandable that most chromosomes contribute to both the DNA located in the internal and peripheral parts of the nuclei, according to the cartoon of Fig. 7.27. This is possible because the GC-richest and the GC-poorest bands are largely located distally and proximally, respectively (and are, therefore, only rarely adjacent in human chromosomes), and because of the large extension of interphase chromosomes. The observation that many chromosomes, especially the large ones, are spread in the nucleus from the periphery to the interior, even if in different proportions (Boyle et al., 2001), supports a compositional polarity of the interphase chromosomes.

We also showed that the two subsets of bands investigated here not only have different locations in the nucleus, but also correspond to different chromatin conformations, the H3⁺ band DNA being remarkably more relaxed compared to the L1⁺ band DNA. This is understandable if we consider the special properties of H3 isochores compared to L1 isochores. In fact, the more open chromatin of the H3⁺ band DNA corresponds to the highest concentration of genes and to the highest level of transcriptional activity. This

![Figure 7.26. Chromosome distribution of the H3⁺ and L1⁺ bands. Ideogram of human chromosomes at 850-band resolution showing the position of the L1⁺ (blue) and the H3⁺ (red) bands as previously identified on the basis of the concentration of L1 and H3 isochores, respectively. This figure shows that the gene-richest, GC-richest, H3⁺ bands are prevalently located distally on each chromosome arm and are generally not adjacent to the gene-poorest, GC-poorest, bands, which are prevalently located more proximally. The intermediate bands (H3⁻, L1⁻) are left uncolored in order to emphasize this spacing. (From Saccone et al., 2002)](image-url)
Figure 7.27. A cartoon showing that the centromeric proximal regions of chromosomes tend to be present in compact chromatin structures at the periphery of the interphase nucleus (blue blocks), whereas the distal regions are in an open chromatin structure (red filaments) in the center of the nucleus. Only two chromosomes are depicted. In this simplified drawing entire chromosome arms are represented as completely open. More realistically, the open chromatin loops should concern a number of regions from each chromosome arm.

Finding is in general agreement with other data showing that chromatin belonging to G and R bands of mitotic chromosomes is more dense and more open, respectively (Croft et al., 1999; Yokota et al., 1997), and that highly expressed sequences extend outside the chromosome territory (Volpi et al., 2000). Incidentally, the latter observation concerned one of the chromosomal regions, 6p21, which was investigated here.

It should be stressed that the two “gene spaces” previously described, the gene-rich “genome core” and the gene-poor “empty quarter” correspond to the genome compartments located in the interior and at the periphery of the nucleus, respectively. This supports a functional compartmentalization of the nucleus related to nuclear architecture (Strouboulis and Wolfe, 1996). The functional compartmentalization of the cell nucleus was also demonstrated by other findings, such as the different replication timing of DNA located in the internal and peripheral part of the nucleus (Sadoni et al., 1999; Ferreira et al., 1997).

Basically, the chicken genome showed the same compositional properties observed for the human genome, both at the chromosomal and the nuclear level, the only difference being the large number of GC-richest bands that are present on microchromosomes (see Andreozzi et al., 2001). In chicken, the large majority of microchromosome DNA only contribute to the internal part of the nucleus, like the small GC-richest human chromosomes, whereas the large chromosomes, comprising both GC-richest and GC-poorest bands, contribute to both the interior and the periphery of the nucleus.

The different spatial distribution of genes in the nucleus and the different chromatin compaction are apparently general for all vertebrates, as expected from our previous comparative investigations, and as shown by results on Rana esculenta (Fig. 7.28).
Figure 7.28. Nuclear distribution of chromosomal regions of *Rana esculenta*. GC-poor and GC-rich chicken DNA fractions were labelled with digoxigenin, hybridized on interphase nuclei, and detected with fluorescein-conjugated antibodies (green signals). Nuclei were stained with propidium (red color). **A.** GC-poor isochores are predominant at the nuclear periphery. **B.** GC-rich isochores are predominant at internal locations. The yellow color is due to overlapping red and green signals. (From Federico et al., 2003).

Further studies should now probe in more detail the distribution of H3+ and L1+ isochores in chromosome territories (see Cremer and Cremer, 2001, for a review). The observations of Tajbakhsh et al. (2002) indicated that the former were localized in all subvolumes of the territories at similar frequency, while the latter were more to the interior of the same territories.

Finally, it should be mentioned that a polymer model for the structural organization of chromatin loops in interphase chromosomes was developed by Ostashevsky (1998) on the basis of isochores.
Part 8
The organization of plant genomes
CHAPTER 1

The organization of the nuclear genome of plants

When we started our investigations on the nuclear genomes of plants in 1986, the available information essentially concerned the buoyant densities, methylation and reassociation kinetics for a small number of species (Ingle et al., 1973; Shapiro, 1976; Sorenson, 1984). As far as primary structures were concerned, data were available for less than 200 coding sequences from all plants. So little being known about plant genomes, what we did was to apply to the nuclear genomes of plants the compositional approach that we had previously used to investigate the organization of the genomes of vertebrates.

The earliest results (Salinas et al., 1988) indicated that the nuclear genomes of Gramineae exhibited strikingly different compositional patterns compared to those of dicots. Indeed, the compositional distribution of nuclear DNA molecules (in the 50-100 kb size range) from three dicots (pea, sunflower and tobacco) and three Gramineae (maize, rice and wheat) were found to be centered around low (41%) and high (45% for rice, 48% for maize and wheat) GC levels, respectively, and to trail towards even higher GC values in maize and wheat.

When the plant sample explored was expanded, it became clear that monocots other than Gramineae could exhibit intermediate or even GC-poor compositional patterns, whereas the dicots explored were all GC-poor, with the possible exception of Oenothera, which showed, next to a very GC-poor component, several GC-rich DNA components (Matassi et al., 1989; see Fig. 8.1).

On the other hand, the compositional distribution of coding sequences from several orders of dicots was found to be narrow, symmetrical and centered around 46% GC, that from monocots (essentially the Gramineae barley, maize and wheat) to be broad, asymmetrical and characterized by an upward trend towards high GC values, with the majority of sequences between 60% and 70% GC. Introns exhibited a similar compositional distribution, but remarkably lower GC levels, compared to the exons from the same genes (see below). The compositional differences in coding sequences between Gramineae and dicots were confirmed by later work on larger data sets (Carels et al., 1998; see Fig. 8.2).

When orthologous coding sequences were compared, points from dicots or from Gramineae, respectively, fell on the diagonal line indicating compositional identity, whereas this was not true when comparison were made between coding sequences from Gramineae (maize) and dicots (Arabidopsis). In this case (Fig. 8.3), both the GC and the GC3 plots showed that, while maize values were close to the diagonal in the low GC or GC3 range (indicating similarity with Arabidopsis values), this was not true for the high GC and GC3 range in which maize values were much higher, yet linearly correlated, with those of Arabidopsis.

In other words, the differences between coding sequences of Gramineae and dicots were confirmed by showing that the compositional distribution of homologous coding sequences from several orders of dicots and from Gramineae mimick the compositional
distributions previously seen for coding sequences in general (Salinas et al., 1988). Indeed, most coding sequences from *Gramineae* were much higher in GC than the homologous sequences of the dicots explored. These differences were even stronger for third codon positions (see above) and led to a striking codon usage for many coding sequences in the case of *Gramineae*.

Introns exhibited a similar compositional distribution, but lower GC levels, compared to exons from the same genes (Salinas et al., 1988; Carels and Bernardi, 2000; see Fig. 8.4). The compositional difference between introns and exons was much greater in plants than in vertebrates. Again, an example of this situation from later work (Carels and Bernardi, 2000) is shown in Fig. 8.5.
Figure 8.2. Distributions of coding sequences A from dicots and B from Gramineae according to the GC₁, GC₂, GC₃ and GC₄ levels, namely the GC levels of first, second and third codon position and of the whole coding sequences, respectively. (From Carels et al., 1998).
Figure 8.3. GC₁, GC₂, GC₃ and GC₄ (GC of coding sequences) of A dicots, B Gramineae, and C maize (ordinates) are plotted against the corresponding values of their homologs from A tomato, B maize, or C Arabidopsis (abscissas). Orthogonal regression lines (solid), diagonals (dashed), and correlation coefficients (R) are shown. (From Carels et al., 1998).
Figure 8.4. Compositional distribution of coding sequences, concatenated introns and genes in maize and Arabidopsis. N is the number of genes investigated (From Carels and Bernardi, 2000)

Figure 8.5. Compositional profiles of the ATPase gene of Arabidopsis (A), ATHATP and of its maize homolog, FMU09989 (B). 5' and 3' flanking sequences are 213 and 260 bp long in the case of Arabidopsis and 1,445 and 2,995 bp in the case of maize, respectively. Exon correspondence and positions are shown in C. (From Carels and Bernardi, 2000)
Two classes of genes were identified in three Gramineae, maize, rice, barley, and six dicots, Arabidopsis, soybean, pea, tobacco, tomato, potato (Carels and Bernardi, 2000). One class, the GC-rich class, contained genes with no, or few short introns. In contrast, the GC-poor class contained genes with numerous long introns (Fig. 8.6). The similarity of the properties of each class, as present in the genomes of maize and Arabidopsis, is particularly remarkable in view of the fact that these plants exhibit large differences in genome size (115 Mb for Arabidopsis; 2,500 Mb for maize), average intron size, and DNA base com-

![Figure 8.6. Distributions of intron number, length of concatenated introns and coding sequence length in GC-poor and GC-rich genes of maize and Arabidopsis. N is the number of genes investigated. An extremely small number of very long concatenated introns in the GC-poor genes from Arabidopsis are not represented. (From Carels and Bernardi, 2000).](image)
position. The functional relevance of the two classes of genes is stressed by (i) the conservation in homologous genes from maize and Arabidopsis not only of the number of introns and of their positions, but also of the relative size of concatenated introns (Fig. 8.7); and (ii) the existence of two similar classes of genes in vertebrates; interestingly, the differences in sizes and numbers of introns in genes from the GC-poor and GC-rich classes are much more striking in plants than in vertebrates. The two classes of genes which differ in intron size have been very recently confirmed (Yu et al., 2002) for Arabidopsis and human.

![Figure 8.7. Number of introns and length of concatenated introns in homologous genes of maize and Arabidopsis. GC-rich genes are represented with solid circles and GC-poor genes with open circles. Number close to circles are the number of genes represented. The solid line is the orthologous regression line through the points, and the broken line is the diagonal. (From Carels and Bernardi, 2000).](image)

An important implication of the results of Carels and Bernardi (2000) derives from the observation that the two classes of genes are conserved in two genomes that widely differ in their GC and GC$_3$ ranges. Indeed, these ranges are very narrow in Arabidopsis (see Figs. 8.13 and 8.14) and very broad in maize (see Fig. 8.8). Since monocots split from ancestral dicots (about 200 Mya), this suggests that the two classes of genes preceded in evolution the compositional genome transition that led to the increased GC levels of a large percentage of genes from Gramineae.
CHAPTER 3

Gene distribution in the genomes of plants

3.1. The gene space in the genomes of Gramineae

The first attempt to investigate the gene distribution in plant genomes was made by localizing 23 nuclear genes from three dicotyledons, pea, sunflower and tobacco, and five monocotyledons of the Gramineae family, barley, maize, rice, oat, and wheat, in DNA fractions obtained by preparative centrifugation in Cs$_2$SO$_4$/BAMD density gradients (Montero et al., 1990). Each one of these genes (and many other related genes and pseudogenes) was found to be located in DNA fragments (50-100 kb in size) that were less than 1-2% GC apart from each other in any given genome. This was taken as an indication of the existence of isochores in plant genomes, namely of compositionally fairly homogeneous DNA regions at least 100-200 kb in size. Moreover, the GC levels of the 23 coding sequences studied, of their first, second and third codon positions, and of the corresponding introns were found to be linearly correlated with the GC levels of the ‘isochores’ harbouring those genes. (Coding sequences for seed storage proteins and phytochrome of Gramineae deviated from the compositional correlations just described). Finally, CpG doublets of coding sequences were characterized by a shortage that decreased and vanished with increasing GC levels of the sequences.

The general picture which was emerging from investigations on plant genomes suggested a strong similarity with results previously obtained for vertebrate genomes. In apparent agreement with this conclusion, the maize genome (2,500 Mb), on which initially we focused our work, had a relatively similar compositional distribution of DNA molecules and of coding sequences compared to the human genome (3,200 Mb), suggesting a similar gene distribution.

Further investigations led, however, to a completely different picture as far as maize and other Gramineae were concerned (for Arabidopsis and other plants see Sections 3.3 and 3.4). Indeed, when the gene distribution of maize was studied using a larger number of probes, a most striking result was obtained, in that almost all genes were found in DNA fragments (50-150 kb in size) covering an extremely narrow (1-2%) GC range and only representing 10-20% of the genome (Carels et al., 1995). This gene distribution (Fig. 8.8), defining what we called a gene space, is remarkably different from the gene distribution previously found in the human genome, and deserves several comments. Moreover, storage protein genes were only in part located in the gene space, several of them being comprised in a small zein space slightly lower in GC compared to the gene space already mentioned. The correlations originally observed were, in fact, essentially due to the existence of two groups of genes, the GC-poor genes of dicots and the GC-rich genes of Gramineae.

(i) The gene space may be visualized as a single family of isochores, since the GC range of the long sequences containing protein-encoding genes (except for some of the zein genes) is
Figure 8.8. The CsCl profile of human DNA (A) is compared with that of maize DNA (B). Both profiles were obtained by centrifuging high molecular weight (50-150 kb) DNA preparations in a CsCl analytical density gradient. The GC scale of the maize DNA profile was obtained from HPLC nucleoside analyses (Matassi et al., 1992) of fractions obtained by centrifuging DNA in preparative Cs2SO4 density gradients in the presence of BAMD. The slightly different relationship of the GC scale to the buoyant density scale in A and B is due to the fact that the high methylation level of maize DNA (Matassi et al., 1992) causes a shift of its CsCl profile toward lower buoyant density values. For human DNA, solid, darkly shaded and lightly shaded areas correspond to high, medium and low gene concentration in the H3, H1+H2 and L1+L2 (namely the very GC-rich, GC-nch and GC-poor) isochore families, respectively. For maize, the solid area corresponds to the gene space; zein genes occupy, however, not only the gene space but also DNA compartments (the zein space) corresponding to the lightly shaded area to its left; the open areas represent genome regions where no protein encoding genes were detected. Ribosomal genes were centered at the buoyant densities marked rDNA. The relative number of genes from human (A') and maize (B') are plotted against GC3 of the corresponding coding sequences. The histogram of 2,762 human genes is from GenBank. Maize data are from GenBank and concern 265 genes plus 41 zein genes (shown as shaded bars). The vertical broken lines correlate the gene distributions (lower frames) with the DNA distribution (upper frame). In maize, such vertical lines only bracket the GC3 range of the genes tested. (From Carels et al., 1995).

narrower (1-2% GC) than that exhibited by any of the human isochore families. That the 50-150 kb DNA fragments actually separated by ultracentrifugation derive from larger chromosomal segments, by fragmentation during DNA preparation, is supported by the fact that hybridization detects segments of DNA which are at least up to twice as large as the DNA fragments tested, namely 100-300 kb (see Fig. 1.11). All other sequences forming the maize genome appear to correspond to gene-empty space (with the exception of those containing ribosomal genes and some of the zein genes). This situation is strikingly different from that found in the human genome, where genes are spread over all isochores, which cover a broad GC range, 30% (see Fig. 8.8).

(ii) Even though the original gene sample tested contained only 21 genes (later expanded to 38; Barakat et al., 1997), these were chosen at random, covered almost the whole range of GC3 values of maize genes, and encoded functionally different proteins. In other words,
the sample used was a representative one, and the conclusions drawn from it could be extrapolated to other maize genes, especially since the probes tested detected not only the genes specifically probed, but also other genes from the same families. These related genes, for which we did not have sequence information, were also located in the gene space. In other words, many more genes were located in the gene space than those that were actually probed.

(iii) The finding that most maize genes are concentrated in a gene space representing 10-20% of the genome and covering an extremely small GC range indicates that most repeated sequences were located outside the gene space. Indeed, repeated sequences form at least 80% of the maize genome (Flavell et al., 1974; Hake and Walbot, 1980; slightly lower figures, 64-73%, were proposed more recently by Meyers et al., 2001) as opposed to 40% in the human genome, and cannot physically fit into the gene space, which only represents 10-20% of the genome, even though the latter is known to contain some particular sets of repeated sequences in its intergenic regions (see the following point).

(iv) The gene space in the maize genome appears to correspond to the only genome compartment in which certain mobile sequences can be transposed. For instance, Mutator, Activator and the majority of Cin4 elements are exclusively located in the same isochores as the Adh-1 gene (Capet et al., 1993), i.e. within the gene space. Maize transposable elements appear, therefore, to be located in gene-rich, transcriptionally active, isochores. Interestingly, this is in agreement with previous observations on transcribed proviral sequences that also integrate into transcriptionally active regions of mammalian genomes (see Part 6). It is also in agreement with results on the localization of T-DNA in the rice genome (Barakat et al., 2000; see Chapter 5).

(v) The results presented here for the maize genome are also valid for the genomes of other cereals. This is suggested not only by general considerations, such as the existence of large regions of collinearity among these genomes and homology between coding sequences, but also by results obtained for wheat, rice and barley. In wheat, hybridization

![Figure 8.9. CsCl profiles of the DNAs from maize, rice and barley, as obtained by analytical centrifugation of high molecular weight DNA (50-100 kb) in a CsCl density gradient. The gene spaces are indicated by the solid areas. The compartments containing zen genes in the maize genome reach buoyant densities as low as 1.7010 g/cm³ (vertical line). Almost all the probes tested on DNA fractions from barley and rice were localized in the gene space. RZ166 and BCD348 were, however, localized in fractions having buoyant densities of 1.7010 and 1.7012 g/cm³ (vertical lines), respectively, in the DNAs of rice and barley. The arrows indicate the localization of ribosomal genes. (From Barakat et al., 1997)](image-url)
of all the genes tested (histone H3; rbcS, ribulose-1,5-biphosphate carboxylase; Cab, chlorophyll-a/b binding protein; α-amylase) on DNA fractions yielded results essentially identical to those obtained on maize fractions (Montero et al., 1990). Similar conclusions can be drawn from hybridization experiments using maize gene probes. Moreover, probes for genes of storage proteins (high molecular weight glutenin; α, β-gliadin) hybridized to fractions exhibiting slightly lower buoyant densities compared to other genes (Salinas et al., 1988; Montero et al., 1990), as was the case with zein genes in maize.

In the case of rice and barley, we found (Fig. 8.9) that the distribution of genes in these genomes is basically similar to that of maize in that all genes, except for ribosomal genes and some storage protein genes, were located in gene spaces that (1) cover GC ranges of 0.8%, 1.0% and 1.6% and represent 12%, 17% and 24% of the genome of barley, maize and rice, respectively (see Table 8.1), and largely consist of transposons; (2) have sizes approximately proportional to genome size, suggesting that expansion-contraction phenomena proceed in parallel in the gene space and in the gene-empty regions of the genome; and (3) only hybridize on the gene spaces (and not on the other DNA fractions) of other Gramineae.

(vi) The gene space consists of a number of small compositional compartments located on all chromosomes, as judged from the known chromosomal location of maize genes (Ahn et al., 1993). Most of the loci tested in wheat physically map in the distal region of the chromosomes (Flavell et al., 1993), where most recombination events also occur (Flavell et al., 1993; Gill et al., 1993). In contrast, the pericentromeric region of a rye chromosome (Moore et al., 1993) shows a reduced density in unmethylated Not I sites, which are indicative of CpG islands associated with genes (Antequera and Bird, 1988; Gardiner-Garden et al., 1992). Interestingly, the high gene concentrations and the high recombination levels detected in telomeric regions of the wheat genome are reminiscent of similar findings on the human genome (Saccone et al., 1992, 1993) and may reflect a general situation.

(vii) The sequence of the Adh-1 280 kb region of maize (Fig. 8.10) published by San Miguel et al. (1996) has shown that the composition of this region (47% GC) exactly corresponds to that of the gene space. This region is very largely made up of nested transposons. Indeed, retrotransposition insertions have doubled the size of the maize genome in the last three million years (San Miguel et al., 1998). Presence of species-,
Figure 8.10. Structure of the Adhl-F region of maize, showing identified retrotransposons. Bar patterns correspond to the approximate copy numbers: cross-hatched bars, highly repetitive (thousands of copies); hatched bars, repetitive (hundreds of copies); and open bars, low copy number (< 100 copies). Letters above bars indicate the class of repetitive DNA, as defined by hybridization of each highly repetitive fragment to all fragments in the region. Confirmed elements have been positioned above the DNA into which they have inserted. Curved lines below each element converge at the insertion site. The arrow above each element indicates its orientation. Fragments with lower case a, b, or c designations are components of the fragment with the same number. (From San Miguel et al., 1996).

and family-specific repeats strongly suggest repeated invasions by different retrotransposons over time (Sandhu and Gill, 2002, and references therein). Needless to say, this abundance of retrotransposons in the intergenic sequences of the gene space eliminate any compositional correlation between coding sequences and the surrounding non-coding sequences. This is in sharp contrast with the situation found in Arabidopsis (see following Chapter 3.2).

3.2. Misunderstandings about the gene space of Gramineae

A recent paper (Meyers et al., 2001) claimed that “An analysis of the GC content of the maize genomic library and that of maize ESTs did not support recently published data that the gene space of maize is found within a narrow GC range”. The claim made by the authors is puzzling in that our actual observations of gene-rich fractions are disputed on the basis of “the demonstration that maize coding sequences have highly variable GC content” (something we have known since the work of Salinas et al., 1988), “that the Huck family of retroelements has a markedly different GC content with respect to the rest of the genome” (also a well-known point) and “that Huck is extremely frequent and dispersed in the genome, as well as associated with all other abundant class of repeats. It is therefore, unlikely that its
Figure 8.11. Chromosomal distribution of the 6591 rice EST sites. Genetic (left; cM) and YAC-based physical (right; kb) maps are shown for each chromosome. The solid lines across each chromosome show the positions and regions of YAC contigs on the genetic map. Different colors represent different EST densities observed on the individual contigs. (From Wu et al., 2002).
distribution could justify the observations of the Bernardi group... which is still publishing the conflicting claim that nuclear genes in Gramineae are found in a 1-2% GC range (Barakat et al., 2000).” Since we were unable to convince the authors with what they call “putative gene rich fractions”, we present in Fig. 8.10 the results of San Miguel et al. (1996) which show the situation of transposons in the 280 kb Adh locus of maize and where the GC level is exactly that found (47%) by hybridizing the Adh probe on DNA fractions. One should not forget that the majority of maize transposons are in the 42-47% GC range, with Huck being 60% GC, but only representing about 10% of the maize genome and being concentrated around centromeric regions (Ananiev et al., 1998) which represent at least 30% of maize chromosomes. It is understandable, therefore, that the gene space is about 47% GC, and coincides with the modal buoyant density of maize DNA.

Another objection raised for the gene space by Dubcovsky et al. (2001) and San Miguel et al. (2002) was that BACs (Bacterial Artificial Chromosomes) from barley, rice and wheat do not differ significantly in composition from total DNAs. This is, however, not surprising if one considers that the gene spaces of Gramineae explored so far practically coincide with the modal buoyant density of DNA (see Fig. 8.9).

A similar misunderstanding was apparently due to a superficial reading of our papers. Sasaki et al. (2002) state, for instance, that “Buoyant density experiments have shown that rice genes are localized in (G+C)-rich islands that occupy 24% of the genome (Barakat et al., 1997). When we plotted the average G+C values against chromosomal position in chromosome 1, however, we did not detect any CpG islands indicating a neutral nucleotide distribution”. As shown in Fig. 8.9, the “islands” (to use the authors’ nomenclature) forming 21% of the genome (see Table 8.1) are very close to the modal GC level of rice DNA and are not GC-rich. It is, therefore, not surprising that G+C-rich islands (mistakenly called CpG islands by the authors) were not found. Incidentally, the assessment of genes by Sasaki et al. (2002) presents the same problem as that of Yu et al. (2002; see Fig. 10.13), namely an extraordinary abundance of imaginary genes.

In fact, our original observation of a gene space in maize (Carels et al., 1995) was followed by other findings pointing towards the existence of gene-rich regions in cereals. Studies on the genome organization of wheat suggested that more than 85% of the genes are present in less than 10% of the chromosomal regions (Gill et al., 1996a,b; Sandhu and Gill, 2002). Results by other authors (Panstruga et al., 1998; Feuillet and Keller, 1999) point in the same direction. The most recent and comprehensive evidence comes from the mapping of over 6500 Expressed Sequence Tag (EST) sites in rice (Wu et al., 2002; see Fig. 8.11).

3.3. The gene space of other plants

In order to decide whether the genome organization and gene distribution found in Gramineae are specific to this family of plants or related to genome size, we have studied the gene distribution in two dicots, pea and tomato, and one non-graminaceous monocot, date-palm (Barakat et al., 1999). The genome of these plants are 5,000, 1,000 and 250 Mb in size, respectively. While in pea most genes are localized in a gene space formed by long gene clusters that cover a 1% GC range and represent about 20% of the genome, in tomato
the vast majority of genes are spread over the genome, covering a 5% GC range which represent 60% of the genome. In date-palm, genes are distributed in DNA fractions covering a 5% GC range and representing 40% of the genome (Fig. 8.12). We conclude that the concentration of genes in a narrow gene space is not an exclusive property of *Gramineae*, because it is also found in a dicot, pea, whereas tomato and date-palm show gene distributions which are closer to that of *Arabidopsis* (see following chapter) than to that of maize.

3.4. Distribution of genes in the genome of Arabidopsis

When we analyzed (Barakat et al., 1998) the small (ca. 115 Mb) nuclear genome of *Arabidopsis thaliana*, which comprises more than 25,000 genes, we could show that its organization is drastically different from that of the genome of *Gramineae* (see Fig. 8.13). Indeed, (i) genes are distributed over about 85% of the main band of DNA in CsCl and cover an 8% GC range; (ii) ORFs are fairly evenly distributed in long (> 50 kb) sequences from GenBank that amount to about 10 Mb; and (iii) the GC levels of protein-coding sequences (and of their third codon positions) are correlated with the GC levels of their flanking sequences (Fig. 8.13).

The recent genome sequence of *Arabidopsis* (Arabidopsis Genome Initiative, 2000) allowed us to construct the corresponding compositional map (Fig. 8.14) which is characterized by only a few moderately GC-rich regions in an overall GC-poor genome. This dicot contains only 20-30% of repeated sequences, half of which are highly repetitive, the other half being moderately repetitive (Meyerowitz, 1992). Moreover, the *Arabidopsis* genome contains very few retrotransposons (the *Ta* and *Athila* families occur in 15 and 150 copies, respectively; Thompson et al., 1996) and is characterized by a very low methylation level (less than 6% of all C are methylated; Leutwiler et al., 1984), the majority of methylation being in centromere-associated tandem repeats, rDNA arrays, and transposons or retrotransposon-derived elements.

Finally, the fact that other members of the *Brassicaceae* family (which diverged very recently from *Arabidopsis*) are endowed with large genomes supports the view that the small genome of *Arabidopsis* is the result of a marked contraction, in which all (or most) of
the large gene-empty regions separating the gene clusters disappeared, as also did most transposons located in intergenic sequences. The contraction undergone by the *Arabidopsis* genome is reminiscent of that shown by the genomes of *Tetraodontiformes*. In this case, highly and moderately repeated sequences were greatly reduced (Pizon et al., 1984) compared with the genomes of its ancestral order *Perciformes* and introns also underwent a contraction. Similar phenomena have been observed in the *Arabidopsis* introns (Carels and Bernardi, 2000).
3.5. A comparison of the genomes of Arabidopsis and Gramineae

Two major differences distinguish the genome of *Arabidopsis* from the genomes of *Gramineae*. The first one concerns the nature of intergenic sequences. In the genomes of *Gramineae*, intergenic sequences within the gene clusters that form the gene space (see Fig. 8.16) are compositionally rather homogeneous, and largely formed by transposons (San Miguel et al., 1996). Indeed, all transposons investigated so far, which include *Mu, Ac* and the majority of *Cin4* elements, are exclusively located within the gene space. This is
Figure 8.15  A. Localization of T-DNA integration in the genome of transformed *Arabidopsis*. Transgenic plants were obtained by transformation with *Agrobacterium* plasmids, which contained a *uidA* (GUS) reporter gene. A GUS probe was hybridized on Cs$_2$SO$_4$/BAMD DNA fractions restricted with *Hind*III, an enzyme not cutting the GUS sequence. *Arabidopsis* DNA was pooled from 30 plants. P, pellet; 1-12, DNA fractions. B. Hybridization of GUS sequence on Cs$_2$SO$_4$/BAMD fractions from rice. 1-12, DNA fractions. (From Barakat et al., 2000).

Figure 8.16. A scheme of genome organization and gene distribution in plant genomes. A. In the large genomes of *Gramineae*, genes (large vertical boxes) are present in long gene clusters, which are separated from each other by gene-empty regions formed by repeated sequences (thick solid line). The ensemble of gene clusters forms the gene space. The intergenic sequences are compositionally very homogeneous because largely formed by transposons (small horizontal boxes in the intergenic sequences). B. The small genome of *Arabidopsis* essentially differs from the genomes of *Gramineae* because of (i) the disappearance (or very strong reduction) of gene-empty regions; (ii) the practical absence of transposons in intergenic sequences; and (iii) the higher gene density. (From Barakat et al., 1998).
also true for the other transposons localized near the \textit{Adh-I} gene by San Miguel et al. (1996). Maize transposable elements appear, therefore, to be located in gene-rich, transcriptionally active regions. As a consequence of the massive integration of transposons into the gene space of \textit{Gramineae}, compositional correlations between coding and flanking sequences are barely detectable. In contrast, in the \textit{Arabidopsis} genome, which comprises a negligible amount of transposons (Thompson et al., 1996), intergenic sequences are compositionally well correlated with coding sequences (see Fig. 8.13).

Interestingly, when we investigated the integration of a T-DNA (transferred DNA from \textit{Agrobacterium} plasmids) in the genomes of \textit{Arabidopsis} and rice, we found different patterns of integration, which are correlated with the different gene distributions (Fig. 8.15). While T-DNA integrates essentially everywhere in the \textit{Arabidopsis} genome, integration in the rice genome was detected only in the gene space, namely in the gene-rich, transcriptionally active, regions (Barakat et al., 2000). This suggests a correlation between integration and transcriptional activity, in agreement with previous observations on transcribed proviral sequences that also integrate into transcriptionally active regions of mammalian genomes (see Part 6).

The second difference concerns gene distribution. As shown in Fig. 8.16 in the genomes of \textit{Gramineae}, genes are present in long gene clusters separated by long gene-empty regions. The ensemble of long gene clusters forms the gene space which corresponds to 12-24\% of the large genomes studied so far. Interestingly, this organization is basically the same over a more than 10-fold range of genome sizes, adding one more basic property to those shared by the genomes of \textit{Gramineae}, like the collinearity of genetic maps and the compositional correlation among orthologous genes. In contrast, in \textit{Arabidopsis}, genes are fairly evenly distributed over regions amounting to about 85\% of the genome, whereas

![Figure 8.17](image_url)

Figure 8.17. Scheme of the DNA and gene distribution in the amphidiploid genome of tobacco. The two Gaussian curves (shifted apart from each other for the sake of clarity) represent the DNA distributions of \textit{N. tomentosiformis} and \textit{N. sylvestris}, respectively; the envelope of the two curves (broken line) represents that of \textit{N. tabacum}. Black boxes represent genes (NAR=nitrate reductase; POD=lignin forming peroxidase) localized in the GC-rich and GC-poor isochores of \textit{N. tomentosiformis}. Gray boxes represent the distribution of Tnt1, which may be accounted for by its mobility. (From Matassi et al., 1991).
gene-empty regions are greatly reduced and may largely consist of the repeated sequences localized in centromeres and telomeres.

3.6. The bimodal gene distribution in the tobacco genome

When we studied the compositional distribution of six genes (or small multigene families) and one family of transposable elements, Tnt1, in DNA fractions from tobacco (Nicotiana tabacum) separated according to base composition, we could show that gene distribution is bimodal and that such bimodality is due to the different base composition of the two parental genomes of tobacco (N. sylvestris and N. tomentosiformis) and to the different parental origin of the genes tested (Matassi et al., 1991).

These results indicate a conservation of the compositional patterns of the two parental genomes, as well as a conservation of the gene localization. (Fig. 8.17) Moreover, they suggest the absence of any extensive recombination between two genomes which have been in the same nucleus for a time which has been estimated to be less than 6 million years, whereas the two parental genomes have an estimated life of about 75 million years (Okamura and Goldberg, 1985).

In conclusion, gene distribution is bimodal in the tobacco genome and this bimodality is due to the amphidiploid nature of this genome. The lack of recombination seen in the tobacco genome fits with the concept of nuclear architecture, namely with the concept of spatial ordering of chromosomes such as that seen in sexual hybrids between barley and rye (see Heslop-Harrison and Bennet, 1990, for a review). Interestingly, a spatial chromosome distribution was also reported in the case of hybrids between N. plumbaginifolia and N. sylvestris. Telocentric chromosomes of N. plumbaginifolia were positioned predominantly at the periphery of metaphase plates, whereas N. sylvestris chromosomes occupied the center (Gleba et al., 1987). Another consideration derived from the tobacco results is that the mixing of two genomes is less obvious than proposed by some authors (see for example, Margulis and Sagan, 2002).

3.7. Methylation patterns in the nuclear genomes of plants

Methylation was investigated in compositional fractions of nuclear DNA preparations (50-100 kb in size) from five plants (onion, maize, rye, pea and tobacco), and was found to increase from GC-poor to GC-rich fractions. This methylation gradient showed different patterns in different plants and appears, therefore, to represent a novel, characteristic genome feature which concerns the noncoding, intergenic sequences that make up the bulk of the plant genomes investigated and mainly consist of repetitive sequences. Our data, as well as data from literature, indicate that both mC and mC/mC+C increase with increasing GC of plant genomes (Fig. 8.18).
Figure 8.18. Plot of mC against the GC levels of plant DNAs. Plants investigated by Matassi et al. (1992) as well as Arabidopsis are indicated. Data for other plants are from references given in the original paper. The correlation coefficient was 0.61 (p=0.0028). (From Matassi et al., 1992).
Part 9

The compositional patterns of the genomes of invertebrates, unicellular eukaryotes and prokaryotes
CHAPTER 1

The genome of a Urochordate,
Ciona intestinalis

The results obtained on the genome of vertebrates prompted investigations on the genome of other eukaryotic genomes. Among those, the genomes of the immediate ancestors of vertebrates, Urochordates and Cephalochordates, were of special interest to us in view of our work having been focused on the organization and evolution of the vertebrate genome. Our knowledge of the genome of Amphioxus, a Cephalochordate, is still too limited to be discussed here. In contrast, we have recently obtained some information about the genome of a Urochordate, Ciona intestinalis, which will be presented in this Chapter.

The CsCl analytical profile of DNA from Ciona is characterized by a main peak with a modal buoyant density of 1.6945 g/cm³, corresponding to 35.16% GC, and by two minor peaks characterized by a lower (1.6707 g/cm³) and a higher (1.7096 g/cm³) buoyant density (Fig. 9.1A). The lighter peak probably is mitochondrial DNA, the heavier peak was demonstrated to correspond to ribosomal DNA by hybridization of an appropriate probe on a preparative CsCl gradient of Ciona DNA. The analytical CsCl profile of the main peak of Fig. 9.1A covers a narrow range. The standard deviation of the CsCl profile (which is the sum of standard deviations due to compositional heterogeneity and Brownian diffusion), was estimated to be 2.36 by fitting a gaussian curve (Fig. 9.1B).

Sixteen single-copy coding sequences were hybridized on DNA fractions from shallow CsCl gradients. GC levels of the fractions corresponding to hybridization peaks are indicated on the analytical CsCl profile in Fig. 9.2A. The GC and GC₃ levels of the coding sequences are shown in Fig. 9.2B. They cover most of the compositional spectrum of the

Figure 9.1. A. Profile of Ciona DNA (molecular size > 100 kb), as obtained by analytical ultracentrifugation to sedimentation equilibrium in a CsCl gradient. The 1.742 g/cm³ peak is the peak of the marker DNA from phage 2C. B. Fit of analytical profile of Ciona DNA (curve) with a gaussian curve (dotted line). This analysis was performed using the MacCurveFit program. The GC scale was obtained from the buoyant density values using the relationship of Schildkraut et al (1962). (From De Luca et al., 2002)
available coding sequences (see below) from *Ciona*.

**Fig. 9.3A** shows a plot of GC$_3$ levels of the sixteen coding sequences analyzed against the GC levels of the corresponding long sequences (GC fractions) in which the coding sequences were localized. The regression equation was used to position the GC$_3$ distribution of the coding sequences available in data bank (about 100 sequences) relative to the CsCl profile (**Fig. 9.3B**).

The local gene concentration was then calculated by dividing each bar value of the GC$_3$ histogram (bins of 5% width) by the corresponding value on the CsCl profile. This approach (Zoubak et al., 1996) estimates gene concentration across the compositional range of the genome and shows that gene density in *Ciona* is only shifted by 1% GC towards higher values from the DNA peak obtained by analytical ultracentrifugation (**Fig. 9.3B**). A very similar result was obtained by hybridizing a cDNA preparation from a mobile larval stage of *Ciona* on DNA fractions from a shallow gradient (**Fig. 9.4**).

**In conclusion,** the genome of *Ciona* is characterized by a very symmetrical compositional distribution of DNA molecules and by a gene distribution which is barely shifted to higher GC values compared to the distribution of DNA. This is a situation which is remarkably different from that found in fishes where both distributions trail towards high GC values. (see **Part 4**). A speculation (modified from De Luca et al., 2002) on the implication of these differences as far as the formation of two gene spaces of vertebrates are concerned follows.

The very first consideration here is to take into account that genome size is 160 Mb in *Ciona* (Satoh et al., 2002) and about 400 Mb in the fishes endowed with the smallest genomes, like *Arothron diadematus* and other fishes of the order *Tetraodontiformes* (e.g., *Fugu rubripes, Tetraodon viridiflormis*; see **Table 4.2**). In the case of *A. diadematus*, the amounts of rapidly and intermediate reassociating repeated sequences are very small (6% and 7%, respectively), in fact the smallest reported so far for a vertebrate genome (Pizon et al., 1984). In other words, even when in its most compact form, a fish genome is more than twice as large as the genome of a urochordate, like *Ciona*. This is a strong indication of a

![Figure 9.2](image)

**Figure 9.2.** A. Localization of the coding sequences used in hybridisation experiments on the DNA fractions from *Ciona*. B. GC and GC$_3$ levels of the coding sequences are listed. (From De Luca et al., 2002).
Figure 9.3. A. Correlation between GC₃ levels of 16 coding sequences (CDS) from Ciona used for hybridization experiments and the GC levels of the DNA fractions in which genes were localized. The range of angles corresponding to a 5% confidence interval, as calculated according to Jolicoeur (1990), is from 80.4° to 84.7°. The slope corresponds to an angle of 82.6°. The corresponding regression line for human coding sequences and the diagonal are also shown by a thick and a thin broken line, respectively. B. Profile of gene concentration in the Ciona genome, as obtained by dividing the relative number of genes in each 5% GC₃ interval of the histogram of gene distribution (yellow bars) by the corresponding relative amount of DNA deduced from the CsCl profile (blue line). The positioning of the GC₃ histogram relative to the CsCl profile is based on the correlation shown in A. (From De Luca et al., 2002).

Figure 9.4. Profile of gene distribution obtained using Ciona cDNA at mobile larval stage as a probe for hybridization (red curve) The histogram shows the DNA distribution obtained from shallow gradient method. 100 µg of DNA were loaded on the gradient. (From De Luca et al., 2002)
genome duplication in the ancestral line leading to fishes, in agreement with the original proposal by Ohno (1970) and some current ideas. In fact, this genome duplication was accompanied by an increase of intergenic sequences, as indicated by the fact that, neglecting cases of later polyploidizations, as in Salmonids, the “average” genome size of fishes is around 1,000 Mb (Bernardi and Bernardi, 1990a,b; see Part 4). This suggests that one possible mechanism for the formation of the two gene spaces in the vertebrate genome was (i) the existence of two compartments characterized by different levels of gene expression in urochordates, a point that can be tested when the annotated sequence of the Ciona genome will become available; (ii) a sequence expansion of intergenic sequences (and introns) in the low-expression compartment which became gene-poor compared to the rest; (iii) the preferential integration in the ancestral genome core of copies of duplicated genes (see Part 6). This proposal is presented in the scheme of Fig. 9.5, which also displays a phylogenetic tree of bilaterian animals.

---

**Figure 9.5.** A. Phylogeny of bilaterian animals. *Ciona intestinalis* is a member of the urochordates, the most primitively branching clade of chordates. Chordates in turn are deuterostomes, one of two great divisions (along with protostomes) of bilaterian animals. (Modified from Delhal et al., 2002). B. A hypothetical scheme of genome evolution from urochordates to fishes, to mammals and birds. In urochordates, DNA and gene distribution are essentially uniform. Two “ancestral gene spaces”, characterized by a low (l) and a high (h) level of gene expression, are however, supposed to exist. The genome duplication event leading to fishes was accompanied by a sequence expansion of l space, with a consequent decrease of gene concentration in that gene space, and by a preferential insertion of duplicated genes in the “ancestral genome core” which was characterized by a slightly higher GC level (indicated by the pink color). This “ancestral genome core” underwent a further stronger increase in GC indicated by the red color to become the ‘genome core’ of mammals and birds. (Modified from De Luca et al., 2002).
CHAPTER 2

The genome of *Drosophila melanogaster*

Since our investigations on the *Drosophila melanogaster* genome (Jabbari and Bernardi, 2000) were done before the complete sequence became available (Adams et al., 2000), it was necessary for us to confirm that the sequence sample that we used (1,095 DNA fragments > 50 kb in size) was a good representative of the whole genome. It was shown, therefore, in a study of the compositional distribution of large DNA segments, that the GC histogram of *Drosophila* DNA sequences larger than 50 kb had a good fit to the CsCl profile (Fig. 9.6). The latter was obtained using *Drosophila* DNA having approximately the same molecular weight as the large sequences. Fig. 9.6 also shows the same comparison for the human genome (5,456 DNA fragments of > 50 kb).

The compositional homogeneity of large DNA segments was investigated by plotting the GC levels of 50 and 100 kb segments from both *Drosophila* and human DNAs against those of the following (contiguous) segments (Fig. 9.7). Slopes close to unity were found in all cases. Correlation coefficients for the two sizes were 0.71 and 0.85 for *Drosophila*, 0.83 and 0.88 for human. Since correlation coefficients for 10 kb and 20 kb segments in *Drosophila* were 0.58 and 0.67, respectively, whereas in human the corresponding values were 0.79 and 0.83, it appears that correlation coefficients for *Drosophila* reach levels close to the

![Figure 9.6](image_url)

Figure 9.6. The GC histogram of large genomic sequences (> 50 kb) from *Drosophila* (A) and human (B) are superimposed on the CsCl profiles of the corresponding DNAs, as obtained by analytical ultracentrifugation (from Thiery et al., 1976, and Zoubak et al., 1996, respectively). Buoyant densities were converted to GC levels according to Schuldcurr et al. (1962). (From Jabbari and Bernardi, 2000).
human ones only for the larger window sizes (100 kb), whereas the short-range compositional heterogeneity is higher in Drosophila.

As far as the plots of Fig. 9.7 are concerned, it should be noted that a perfectly homogeneous genome would produce a diagram consisting of a single point located on the diagonal, whereas a correlation coefficient of unity (and a unity slope) would correspond to a set of perfectly homogeneous tracts, *i.e.*, to the compositional identity of all pairs of adjacent segments (points outside the diagonal corresponding to compositional differences), and randomly generated sequences would not exhibit any correlation (not shown). Although data available at the time of the work did not allow using window sizes larger than 100 kb, one could still conclude that DNA segments with compositional homogeneity may reach sizes of at least 200 kb, because of the closeness of GC levels of a number of adjacent 100 kb segments.

The compositional correlations of third codon positions with long (> 50 kb) DNA segments in which the coding sequences were present (Fig. 9.8A) showed a slope of the orthogonal regression line equal to 5.5 (the range corresponding to a 5% error being 5.1–5.9) and a correlation coefficient *r* of 0.42 (*p* < 0.0001). These values are consistent with those of the correlations between GC3 levels and the GC levels of 5' and 3' flanking sequences 20 kb in size. The slope is higher than that observed for human, 4.09 (Fig. 9.8B), yet the difference in angles is small, only 3° (see Clay et al., 1996).

The distribution of GC3 values of Drosophila coding sequences (Fig. 9.8) covers a range

![Figure 9.7](image-url). GC levels of Drosophila and human DNA sequences of 50 and 100 kb are plotted against the GC levels of the following contiguous sequences of the same size. *r* is the correlation coefficient, *n* the sample size and *w* the window size. The diagonal is shown as a broken line. (From Jabbari and Bernardi, 2000).
that is comparable to that covered by human coding sequences, but the distribution profile is much narrower and centered on higher values, the mean GC$_3$ value being 66.2 ± 11.6.

The regression equation of Fig. 9.8A (upper panel) was used to position the GC$_3$ distribution relative to the CsCl profile (Fig. 9.8B, upper panel). The local gene concentration was then calculated by dividing each bar value of the GC$_3$ histogram by the corresponding value on the CsCl profile. This approach (Zoubak et al., 1996) estimates gene concentrations across the compositional range of the genome, and shows that gene density in *Drosophila* is characterized by low levels in the GC-poorest part of the genome (and, expectedly, by no genes in the GC-poor satellite DNA) and by a linear increase over the rest of the genome. Fig. 9.8B shows that the profiles of gene distribution are comparable in the human and *Drosophila* genomes, in that gene concentration parallels GC concentration in both cases. However, gene concentration showed a 7-fold difference between the GC-poorest and the GC-richest regions of the *Drosophila* genome versus a 20-fold factor in the case of the human genome. A still lower, 3-fold, difference was found in the genome of *Anopheles gambiae*, which, incidentally, is 45% GC (and not 35% GC, as claimed by Holt et al., 2002) on the average (Jabbari and Bernardi 2003b). A factor that could conceivably limit the accuracy of gene distribution estimates in the *Drosophila* genome is the lower correlation coefficient of the plot of GC$_3$ vs. GC of large sequences, compared to the corresponding human plot. However, the error in the slope of the regression lines is small.
Fig. 9.9 shows that the compositional patterns of chromosomes from *Drosophila* are characterized by a predominance of relatively GC-rich structures (corresponding to H1 isochores) with very few GC-richer tracts but a relative abundance of GC-poor stretches (corresponding to L2 isochores).

In conclusion, the *Drosophila* genome shows a similarity to the human genome in its gene distribution, although the gradient of gene concentration is less steep in *Drosophila* compared with human. The evolutionary reasons for such similarities will be discussed in Part 12, section 1.2. Moreover, the short range heterogeneity of the *Drosophila* genome is higher than that of human.

Figure 9.9. Compositional profiles of *Drosophila melanogaster* chromosomes. Vertical grey lines correspond to gaps in the sequence. (Modified from Paçes et al., 2004)
The genome of *Caenorhabditis elegans*

The nuclear genome of *Caenorhabditis elegans* is 97 Mb in size and comprises over 19,000 genes rather uniformly distributed over six chromosomes (*Caenorhabditis elegans* Sequencing Consortium, 1998). Compositionally, its DNA is only 36% GC. The compositional patterns of *C. elegans* chromosomes are presented in Fig. 9.10 to show the difference between the autosomes, which are GC-poor and relatively heterogeneous, and the X chromosome, which is GC-poorer and more homogeneous. On the X chromosome, genes appear at a lower density and are more evenly distributed compared to autosomes. In contrast, the frequency of EST matches varies according to their position along the auto-
omes, indicating a clustering of highly expressed genes. Gene density is, therefore, reminiscent of features encountered in other eukaryotic genomes.
The nuclear genome of unicellular eukaryotes

In this chapter, we will concentrate on the compositional patterns and gene distribution found in the very different genomes of unicellular eukaryotes.

(i) The yeast nuclear genome was the first eukaryotic genome to be fully sequenced (see Goffeau et al., 1996). This genome is organized in 16 chromosomes and its total size is only 12 Mb, comprising about 6,000 genes. Compositionally, yeast chromosomes are remarkably homogeneous and GC-poor (Fig. 9.11). Some yeast chromosomes consist, however, of alternating large domains of GC-rich, and GC-poor DNA (Karlin et al., 1993; Sharp and Lloyd, 1993; Dujon, 1996), generally correlating with the variation in gene density along these chromosomes. In the case of chromosome III, it has been demonstrated that the periodicity in base composition is paralleled by a variation in recombination frequency along the chromosome arms, with the GC-rich peaks coinciding with regions of high recombination in the middle of each arm of the chromosome, and AT-rich troughs coinciding with the recombination-poor centromeric and telomeric sequences. Zenzvir et al. (1992) have demonstrated that the relative incidence of double-strand breaks, which are thought to initiate genetic recombination in yeast (De Massy et al., 1995), correlates directly with the GC-rich regions of this chromosome.

(ii) Trypanosomes (Class Zoomastigophora, Order Kinetoplastida) are very ancient flagellated protozoa (Sogin et al., 1989), whose extant species are only known as parasites. In the case of Trypanosoma brucei, the haploid genome consists of three size classes of chromosomes: 11 megabase-sized (ranging from 1 to 6 Mb), plus 50 to 100 minichromosomes (25–100 kb) and intermediate (100–800 kb) chromosomes, totalizing approximately 78 Mb (Melville et al., 2000).

High molecular weight DNA preparations from T. brucei and T. equiperdum, two closely related trypanosomes (Haag et al., 1998; Stevens et al., 1999), were fractionated by preparative centrifugation in Cs2SO4/BAMD density gradient. Analytical centrifugation in CsCl of the DNA fractions so obtained showed that both DNAs had a bimodal distribution with two major peaks banding at 1.702–1.703 and 1.707–1.708 g/cm^3 and representing 1/3 and 2/3 of total DNA, respectively (Fig. 9.12). A number of minor components, corresponding to satellite DNAs and possibly to ribosomal DNA (in the case of the 1.718 g/cm^3 DNA component of fraction 11 of T. brucei) were also detected. These results (Isacchi et al., 1993) indicate that a compositional compartmentalization is not only found in the genome of vertebrates, but also in those of protozoa such as Trypanosomes.

The overall GC range (35–55%) covered by large DNA fragments from these two genomes is very wide. The bimodality of the CsCl profile is very striking in that both components correspond, at least very largely, to protein-encoding DNA. This is indicated by three lines of evidence: (1) hybridization experiments have located ESAG (Expression Site Associated Genes) in the GC-poor compartment and housekeeping genes in the GC-rich compartment of T. brucei; (2) third codon positions of T. brucei display a large
Figure 9.11. Compositional patterns of yeast chromosomes. (Modified from Páez et al., 2004).
standard deviation (10%) and values from 20% to 97%, and the ESAG genes have a lower GC content compared to housekeeping genes (Michels, 1987; Musto et al., 1995); (3) satellite and middle repetitive DNA are scarce in the T. brucei genome.

We have also investigated the compositional distributions of exons and their codon positions, as well as the codon usage and amino acid composition of the nuclear genomes of the African and American trypanosomes T. brucei and T. cruzi. Very large differences between the two species were found in all the properties investigated. The most striking differences concern the compositional distributions of third codon positions and the extremely large nucleotide divergence of third codon positions for homologous genes encoding proteins that are highly conserved in their amino acid sequences. Moreover, when coding sequences from each species were divided into two groups according to the GC levels in third codon positions, very different codon usages and amino acid compositions were found (Musto et al., 1994).

In conclusion, the compositional patterns of the Trypanosome genomes indicate that compartmentalization is a phylogenetically very widespread situation in eukaryotes, as already predicted (Bernardi et al., 1985b).

(iii) Plasmodia. A compositional compartmentalization was demonstrated in the nuclear DNA of Plasmodium cynomolgi, which consists of isochores likely to average 100 kb (McCutchan et al., 1988).

A striking feature of P. falciparum, the unicellular parasite responsible for the most virulent and widespread form of human malaria, is that it hosts the GC-poorest (22% GC) nuclear genome known so far (Pollak et al., 1982, McCutchan et al., 1984). This genome, which only comprises 3 Mb of DNA (Weber, 1988) organized in 14 chromosomes (Kemp et al., 1987; Wellems et al., 1987), is, therefore, an excellent model for studying compositional constraints and their effects.

The analysis of sequence data has provided useful information about the genes encoded in the nuclear genome of P. falciparum. The most relevant features are the following: (i) the coding strand is purine-rich; (ii) A is predominant in all codon positions; (iii) the third codon positions are extremely GC-poor and, as a consequence, codon usage is strongly biased (Weber, 1988; Hyde and Sims, 1987).

In an analysis of the nuclear coding sequences from P. falciparum, comprising 175 kb (Musto et al., 1995), we found that the trends described previously with more limited data sets are still valid. We then tried to understand whether the biases already noted are species-specific or determined by the composition of the genome, and whether they are different for sequences that display different expression levels and certainly are under different evolutionary pressures, like housekeeping and antigen genes (see also Part 11).

In the case of the P. falciparum genes that we analyzed here, we noted that the biases in base composition of different codon positions, amino acid frequencies and codon preferences are almost identical in housekeeping and antigen sequences. If we take into account that both the level of expression and the evolutionary constraints over these sequences are most unlikely to be the same, it is reasonable to conclude (i) that the extremely biased composition of the genome (Pollak et al., 1982; McCutchan et al., 1984) is the major factor in determining codon preferences and amino acid frequencies; and (ii) that the compositional constraints (Bernardi and Bernardi, 1986a) operate in the same direction over all the translated sequences and their codon positions (Musto et al., 1995, 1997, 1999).
Figure 9.12. A. Analytical CsCl profiles of unfractionated (total) nuclear DNAs from *T. brucei* and *T. equiperdum* and of the compositional fractions obtained by preparative centrifugation in a Cs₂SO₄/BAMD gradient. B. Histograms displaying the relative amounts of the major DNA components of *T. brucei* and *T. equiperdum*, as estimated from the data of A. In *T. equiperdum*, fractions 2 and 3 and fractions 8 and 9, respectively, were pooled because of their very close modal buoyant densities. Likewise, in *T. brucei*, fractions 1, 2 (except for the 1.708 g/cm³ component) and 3, and fractions 10 and 11, respectively, were pooled. (From Isacchi et al., 1993).

Our results further support the hypothesis that among the various factors that certainly influence the architecture of the coding (and non-coding) sequences, the compositional constraints on the genome (or on isochores) are most relevant, as already proposed (Bernardi et al., 1985b).
CHAPTER 5

Compositional heterogeneity in prokaryotic genomes

5.1. CsCl gradient ultracentrifugation and traditional fixed-length window analyses

Almost fifty years ago, Lee et al. (1956) reported that bacterial DNAs cover a very wide spectrum of base compositions, a result confirmed by Belozerski and Spirin (1958). Shortly after Meselson et al. (1957) introduced density gradient centrifugation of DNA as a new experimental approach in molecular biology, Rolfe and Meselson (1959) and Sueoka et al. (1959) discovered that the modal buoyant density of bacterial DNAs was correlated with their average base composition, GC, the molar ratio of guanine+cytosine. At the same time, they realized that bacterial DNAs were characterized by very narrow band profiles compared to mammalian DNA. The difference in heterogeneity was overestimated by the fact that the standard mammalian DNA used at that time, calf thymus DNA, comprises 23% GC-rich satellites (as we showed 20 years later; Macaya et al., 1978). The idea of remarkably homogenous genomes stuck, however, to bacteria until present. A detailed analysis by Bernaola-Galvan et al. (2004), outlined below, refutes a widely accepted idea that bacteria and archaea normally exhibit low GC contrasts. This study extended an early compilation of heterogeneity data from ultracentrifuge work on Clostridium, Haemophilus, Micrococcus, Streptococcus, Bacillus subtilis and Escherichia coli (Cuny et al., 1981).

Fig. 9.13 shows intragenomic GC heterogeneity at different scales for human and four bacterial species, as assessed using the traditional fixed-window approach. The difference between a homogeneous (Chlamydia trachomatis) and a heterogeneous example (Xylella fastidiosa) is striking. Whereas C. trachomatis exhibits only a slightly higher heterogeneity than a random sequence, at all scales from 100 bp to 100 kb, X. fastidiosa approaches the heterogeneity of human DNA at scales around 5 kb.

5.2. Generalized fixed-length window approaches

Measuring compositional heterogeneity via the standard deviation of GC levels among fixed-length segments has two important advantages. First, the behavior of this measure, including anomalies and biases, is well understood and has been extensively studied for more than a century. Second, in the absence of sequence information, CsCl density gradient ultracentrifugation can be used to obtain good estimates of the fixed-length fragment distribution of a genome’s GC levels, and of its standard deviation.

Every measure of genomic heterogeneity has, however, characteristic drawbacks. It is therefore advisable to confirm results obtained via one measure by other measures, and by checking that heterogeneities are not simply due to a few short “outlier” regions. For example, the standard deviation of fixed-length fragments is sensitive to outlier DNA
having GC levels that are far from the mean (e.g., satellites or rRNA genes). Another drawback of this measure is seen when genomes are organized into relatively homogeneous regions of very different sizes. A plot of standard deviation vs. window length, as in Fig. 9.13, then only gives an indirect picture. We therefore applied several methods, which allowed us not only to explore different kinds of heterogeneity, but also to confirm the robustness of some conclusions.

The usual fixed-length window approach can be generalized in at least three ways. A first way is to keep fixed windows, but to weight the nucleotides inside each window according to their positions within the window. A second way is to partition the entire sequence, not into fixed-length segments, but into variable-length segments whose lengths are chosen by the GC variation itself. A third way is to find a representation of heterogeneity that separates it into contributions from codon position differences and from heterogeneities above the scale of genes.

Traditional moving-window plots use a box filter: each GC base pair within a fixed length window is counted once. From such windows' GC levels one can recover their standard deviation, the traditional measure of heterogeneity. Alternative filters can be designed, which weight the GC counts according to their position in the window. The weighting can, for example, be made to follow a Gaussian form, in order to emphasize the
center of the window (Fig. 9.14). When window counts are represented by color or brightness rather than by the height of a line plot, a single plot can accommodate the full range of possible window sizes. From such plots one can quickly locate the homogeneous and heterogeneous regions present in a genome at any scale, and assess their stability when the window size changes. Such plots allow quick comparisons between the overall contrast levels present in different genomes, as can be seen in Fig. 9.14. They also localize the regions that cause the contrasts, and indicate if they are confined to a single locus or ubiquitous. *X. fastidiosa* contains one striking, very GC-rich (blue) locus, but also shorter, similarly GC-rich regions throughout its genome. Even after excluding these regions, the overall heterogeneity remains distinctly higher than in *C. trachomatis*. The same conclusion is reached via segmentation analysis (see below).

![Figure 9.14. Fixed-length moving window plots of the GC heterogeneity of two completely sequenced bacterial genomes, *C. trachomatis* and *X. fastidiosa*, using a modification of the traditional box filter. GC counts are given smaller weights when they are located farther from the window’s midpoint, according to a Gaussian weight curve. GC content (expressed as a fraction of 1) is shown as color (dark blue in GC-rich regions). Each horizontal transect is a moving-window plot for a different window size; along the vertical axis, window sizes increase logarithmically from 1 kb to 100 kb. (From Bernaola-Galvan et al., 2004).](image)

### 5.3. Intrinsic segmentation methods

A second alternative to traditional fixed-length window scans is a rigorous partitioning, using criteria that involve the GC variation within the sequence, into relatively homogeneous segments of different sizes. No restrictions are placed on the lengths of the segments. A genome can be partitioned in this way by recursive top-down segmentation (Bernaola-Galvan et al., 1996). First, the entire sequence is searched to find a partition into two parts that would maximize their GC contrast. A chosen significance threshold or stopping criterion then decides if the contrast is large enough to justify the proposed partitioning. The process is iterated until a contrast between two adjacent (sub)segments becomes weaker than the chosen threshold. We applied a recently developed segmentation variant (Oliver et al., 2001, 2002) which incorporates two further key improvements, making it particularly suitable for locating boundaries between fairly homogeneous, isochore-like regions: (i) the cuts on the sequence are made one by one, in a hierarchical way, which ensures the most statistically significant cut at each scale; and (ii) short-scale sequence heterogeneity below 3 kb is filtered out. Such coarse graining of the sequence corresponds
to a homogeneity criterion for mammalian isochores, derived via ultracentrifugation of DNA at different molecular weights. Although the stringency of the stopping criterion is related to the average lengths of the segments, the segment lengths typically span a wide range (see Oliver et al., 2001, 2002, for examples). This approach therefore yields a different picture of a genome’s compositional variation than when fixed segment lengths are used.

Compositional segmentation plots are shown in Fig. 9.15A for eight heterogeneous and two homogeneous genomes. They consolidate the result that was exemplified above for \textit{X.fastidiosa}: in several heterogeneous genomes, the heterogeneity is obviously not just due to one or two small regions having exceptional GC levels, but it is instead a genome-wide phenomenon. The corresponding sequence compositional complexity (SCC) values are also shown (Fig. 9.15B, top scatterplot; block size for permutation test set at 3 kb), as well as those obtained at a higher stringency (bottom scatterplot; block size set at 500 bp). On average, higher stringencies correspond to shorter scales. With respect to the SCC, some species are among the most heterogeneous at both stringencies.

These taxa include \textit{Aeropyrum pernix} (triangle 9), \textit{Halobacterium sp.} (triangle 3), \textit{Bacillus subtilis} (square 37), \textit{Mycoplasma genitalium} (square 61; but not \textit{Mycoplasma pulmonis}, 41 or \textit{Mycoplasma pneumoniae}, 62, which is heterogeneous only at short scales/high stringencies; cf. Kerr et al., 1997), and \textit{Xylella fastidiosa} (square 16). It can be seen that, at the chosen stringencies, the heterogeneity shows no clear relation to the overall GC level of the genome. It can also be seen that neither in archaea nor in bacteria alone is there a general relation to GC level. Furthermore, there is no obvious grouping of hyperthermophiles. These general observations are also valid when the standard deviation of fixed-length windows is used as the heterogeneity measure. No relation was found between genome size and sequence compositional complexity (SCC) at the 3 kb stringency level (a level that corresponds to variation in GC above the typical lengths of genes). For example, \textit{Mesorhizobium loti} has one of the largest genome sizes of the sequenced species (above 7 Mb) and about 8,000 genes (Kaneko et al., 2000), yet it has a relatively low SCC value of 0.8 bits/kb.

Low SCC complexity values were also found in some bacterial intracellular parasites, such as the different strains of \textit{Chlamydia} (0.14–0.35 bits/kb) or the obligate parasite of aphids, \textit{Buchnera} (0.19 bits/kb). The presence of pathogenicity in some genomes (as in \textit{Helicobacter pylori} or \textit{Pseudomonas aeruginosa}) and symbiotic islands (as in \textit{M. loti}) seems not to determine high SCC values. Finally, the high compositional heterogeneity in the archaeon \textit{Aeropyrum pernix} (4.3 bits/kb), and correlations extending until 20 kb may be related to the large number of duplicated ORFs in this genome (Kawarabayasi et al., 1999).

\textbf{In conclusion}, the data presented here show that some prokaryotic genomes are compositionally heterogeneous, as assessed by several criteria at several scales and stringencies. While others are homogeneous. Early data sets consisting of few species had suggested a convex dependence of bacterial heterogeneity on GC level with a maximum near 50% GC (Cuny et al., 1981), yet our larger sample of genomes no longer shows a clear trend of this kind at the scales of stringencies used here: one finds both homogeneous and heterogeneous genomes spanning a wide range of GC levels, and there is no apparent dependence of heterogeneity on GC level. This situation in prokaryotes is different from that found in mammals and birds since at comparable scales, the heterogeneity of isochores increases
Figure 9.15. Compositional segmentations of sequenced prokaryotic genomes and their associated GC heterogeneities as measured by sequence compositional complexity (SCC). A. Segmentations for 8 heterogeneous and 2 homogeneous genomes (bottom panels). The segmentation stringency was determined by a permutation test for 3 kb blocks (see Oliver et al., 2001, for details). B. SCC values for segmentations of sequenced archaeal (triangles) and bacterial (squares) genomes at two different stringencies, corresponding to permutation tests for 3 kb (top panel) and 500 bp blocks (bottom panel). The SCC of a segmented sequence is the difference between the sequence's informational (Shannon) entropy and the sum of all its segments' entropies, weighted by their respective lengths (Roman-Roldan et al., 1998). Horizontal lines in the scatterplots indicate SCC values of human sequences: the very heterogeneous MHC region in chromosome 6, spanning three adjacent isochores; the GC-rich (H) MHC class III isochore; the GC-poor (L) central, classical MHC class II isochore; and the longest (approx. 7 Mb) GC-poor L isochore of chromosome 21. Isochore boundaries are as determined in Oliver et al. (2001). 3 kb is the stringency that reveals the extents of these human isochores. Thus at this stringency they all have SCC values of 0 (top scatterplot), and the SCC levels of many prokaryotic genomes are higher than those of these human isochores. (Modified from Bernaola-Galvan et al., 2004)

...dramatically when GC levels rise from 30% to 50% (Cuny et al., 1981; Bernardi, 2001, and refs. therein).

We have used several measures of GC heterogeneity: the variation among fixed length windows, the sequence compositional complexity at different stringency levels, and the value of the GC correlograms where top and bottom branches meet (not shown here). For
two or more of these measures, we consistently found the following genomes among the
most heterogeneous species: *Xylella fastidiosa, Aeropyrum pernix, Halobacterium sp., Mycoplasma genitalium, Neisseria meningitidis,* the Sakai strain of *Escherichia coli,* and *Vibrio cholerae* (both chromosomes). Among the most homogeneous species we found *Agrobacterium tumefaciens, Chlamydia trachomatis, Rickettsia prowazekii,* and *Campylobacter jejuni.*

Although some genomes' heterogeneities may be partly due to horizontal transfers of DNA from species having different GC levels, such transfers are unlikely to be the only source of heterogeneity in the genomes we have analyzed. For example, the robustly heterogeneous *Mycoplasma genitalium* is often considered to contain almost no recent horizontal acquisitions (Ochman et al., 2000; Kerr et al., 1997). In at least some genomes, therefore, alternative explanations must be considered, such as region-dependent selection on base composition.

### 5.4. Does intragenomic heterogeneity in *E. coli* arise from exogenous or endogenous DNA?

Prokaryotic genomes show a wide range of intragenomic compositional heterogeneities, which can reach values comparable to those of regional heterogeneities in mammalian genomes (see Chapter 1). In fact, we reported, more than 20 years ago, that fish genomes were very often more compositionally homogenous than bacterial genomes in the same size and composition range (Hudson et al., 1980; Fig. 3.10).

The presence of considerable GC contrasts within the apparently endogenous DNA of some genomes, and its absence in others, remains enigmatic. The similarities between the GC distributions of prokaryotes and vertebrates in shape and variety (albeit not in width) raise a question that has received relatively little attention until now: could some of the factors shaping GC contrasts in prokaryotes be similar to those shaping the much more pronounced GC contrasts in vertebrates?

The extent to which lateral transfers can be recognized by compositional contrast is still an unsolved issue. Some authors have, in essence, suggested that it is the recently acquired foreign DNA that occupies most of the GC-poor tail of the *E. coli* genome's asymmetric GC distribution, for example, so that without such foreign DNA the GC distribution would be thinner and almost symmetric. Large GC contrasts in *E. coli* are thus interpreted as likely footprints of relatively recent lateral transfers, and are thought to erode with time. This line of reasoning was partly motivated by early comparisons of sequenced genes in *E. coli* and *Salmonella enterica,* which suggested that genes present in one lineage but absent in the other, *i.e.,* genes that could have arrived by lateral transfer, were frequently aberrant in base composition and/or codon usage. Such results suggested that other compositional anomalies might also indicate a foreign origin of the DNA (see, *e.g.,* Groisman et al., 1993; Lawrence and Ochman, 1997; Ochman et al., 2000).

It is still difficult to evaluate the utility of this working hypothesis as a general principle for prokaryotes, since the currently used compositional criteria often use cutoffs that are obtained on the basis of particular assumptions, rather than on a dataset of lateral transfer genes that have been independently verified for the species studied. An alternative to the
lateral transfer explanation may be selection on base composition in certain regions of prokaryotic genomes, as was pointed out already by Syvanen (1994). More generally, Sueoka (1992) has offered the following perspective: “The (relatively) small intragenomic heterogeneity of G+C contents in unicellular organisms such as E. coli and yeast may be regarded as a simple form of isochores (domains of unique G+C contents). It is a likely possibility that the intragenomic heterogeneity of DNA G+C content is a ubiquitous phenomenon from bacteria to mammals and that the only difference is the extent of heterogeneity between unicellular and multicellular organisms.”

Perhaps the best-known examples of genes that experience compositional constraints in prokaryotes are the ribosomal RNA (rRNA) genes: hyperthermophiles apparently experience strong selective pressure to keep their rRNA genes at high GC levels. The high GC levels ensure the correct secondary structures for the rRNA at high temperatures, and are sometimes maintained in spite of a quite different GC level of the rest of the genome: GC levels of rRNA genes (and tRNA genes) in hyperthermophiles appear to correlate preferentially with temperature, rather than with the overall genomic GC content (Galtier and Lobry, 1997; Hurst and Merchant, 2001); see, however, note added in proof in Part 12, Chapter 5). Exceptional compositions of the region surrounding the origin or terminus of replication might also reflect constraints, e.g. temperature-dependent constraints for structures that are needed for efficient functioning of the ori sequence. Constraints of this type have been shown experimentally for ori sequences in bacteriophage G4 and in the mitochondrial genome of yeast (Goursot et al., 1988, and refs. therein). There appears to be no reason, a priori, why GC constraints should not also conserve the GC content of some regions containing protein-coding genes.

5.5. Inter- and intra-genomic GC distributions

Datasets collected during the past 50 years consistently show a wide range of genomic GC contents among prokaryotic species and genera, although they do not yet suggest a particular shape of the intergenomic GC distribution. In view of the many and diverse prokaryotic taxa that are still being discovered and the strong biases in currently available samples (Hugenholtz, 2002), it seems that it is still too early for even taxonomically filtered subsamples of the available genomic GC levels to yield a representative distribution of bacteria or archaea.

Well before the issue of frequent lateral transfers arose, ultracentrifugation results on bacterial GC heterogeneity (Sueoka, 1959; Guild, 1963; Yamagishi, 1974) had indicated that long regions of distinct base composition should exist within the genomes of some bacteria. The existence of such regions has been confirmed by segmentation analyses of GC contrasts in sequenced prokaryotic genomes (Bernaola-Galvan et al., 2004). In some prokaryotic genomes, one finds compositional mosaicism, while in others, such as M. genitalium, the variation around the genome resembles a wave rather than a mosaic. But the reasons for the diverse types of systematic, large-scale variation remain an open question. The relative contributions of horizontal transfer from chromosomes of distantly related species, of phages, of recombinogenicity and of selection on base composition, for example, are still largely unresolved.
Part 11

The compositional evolution of vertebrate genomes
Two modes of evolution in vertebrates

From a compositional viewpoint, vertebrate genomes evolve, as a general rule, according to a conservative mode: nucleotide substitutions do occur, but the base composition of isochores, and of coding sequences in particular, does not change. In other words, evolution essentially takes place in a condition of compositional equilibrium. Alternative views are discussed in Part 12, Chapters 5 and 6. As it will be shown in the following Chapter 2, this conservative mode is far from obvious, because of the specific compositional patterns characterizing vertebrate genomes and because of the non-randomness of the mutational input.

Under some circumstances (to be discussed in Part 12), vertebrate genomes evolve, however, according to a transitional, or shifting, mode, in which nucleotide substitutions are accompanied by compositional changes. Fig. 11.1 presents a scheme of the conservative mode of evolution and of the three kinds of compositional transitions, or shifts, that we have observed. In Fig. 11.1 the distribution of DNA molecules is approximated by the

Figure 11.1 A schematic representation of A the “major”, B the “minor”, and C the “whole-genome”, or “horizontal”, compositional transitions, or shifts, in the genomes of vertebrates. This scheme displays the CsCl profiles (which are good approximations of the compositional distributions of DNA molecules) along with the changes that follow the compositional shift under consideration (broken lines). A refers to the transitions between cold- and warm-blooded vertebrates; B to the transition between the general mammalian pattern and the murid pattern; C to the transitions among cold-blooded vertebrates; D to the conservative mode of evolution. (Modified from Bernardi et al., 1997).
CsCl profiles. In the transitional, or shifting mode, one part of the DNA molecules changes its composition. Two *intragenomic shifts* have been recognized. The **major shifts** (Fig. 11.1A) concern the changes undergone by the gene-dense part of the vertebrate genome (as well as by the coding sequences embedded in them) at the transitions between reptiles and mammals or between reptiles and birds. The **minor shift** (Fig. 11.1B) concerns the compositional changes undergone by the GC-poorest and the GC-richest isochores at the emergence of murids. In this case, the two extreme compositions converge towards average values. In contrast, the **whole genome or horizontal shifts** (Fig. 11.1C) concern the totality of the genome which becomes GC-poorer or GC-richer, without any change in its compositional pattern. Finally, in the conservative mode (which precedes and follows the shifts), no change takes place (Fig. 11.1D).

One should recall here that the compositional patterns of DNA molecules are correlated with the compositional patterns of coding sequences (see Part 3, Chapter 2, Figs. 3.17 and 3.19). One could, therefore, schematically represent the compositional modes of evolution using coding sequence patterns instead of the patterns of DNA molecules.
CHAPTER 2

The maintenance of compositional patterns

2.1. The maintenance of the compositional patterns of warm-blooded vertebrates

The existence of a compositionally conservative mode of evolution of warm-blooded vertebrates was initially suggested by the similarity of their isochore patterns (which may, however, undergo very limited “whole-genome shifts”, to be discussed in Chapter 5), and by the similarity of their coding sequence patterns. Another line of evidence was provided by \( GC_3 \), \( G_3 \) and \( C_3 \) plots concerning orthologous genes of species from different mammalian orders. Fig. 11.2 shows \( G_3 \) and \( C_3 \) plots for human and calf genes (artiodactyls share with primates the “general mammalian pattern”, the most widespread mammalian pattern, as defined by the CsCl profiles; see Part 4).

Fig. 11.2 shows that the regression lines of the human/calf plots pass through the origin and are characterized by unity slopes and very high correlation coefficients, 0.88–0.89. In other words, average \( G_3 \) and \( C_3 \) values of orthologous genes from human and calf are very close to each other over the entire \( GC_3 \) range, including \( GC_3 \) values up to 90–100%. Now, extant placental mammalian orders diverged some 100 Mya (million years ago; a very approximate figure based on molecular data; Murphy et al., 2001; Eizirik et al., 2001; Hedges and Kumar, 1999) from a common ancestor according to a star-like phylogeny (see Fig. 11.3). In other words, they evolved independently of each other for about 100 Myrs. One should conclude, therefore, that the very many mutations that took place during this long time interval led to no change or only to small non-directional changes in the com-

![Figure 11.2. Correlation between \( G_3 \) and \( C_3 \) values of orthologous genes from human and calf. The orthogonal regression lines coincide with the diagonals. The equations of the regression lines and the correlation coefficients are shown. \( N \) is the number of gene pairs explored. (From Bernardi, 2000b).](image-url)
positional patterns of both DNA and coding sequences. Because of the linear correlations between GC\textsubscript{3} and isochore GC, these plots confirm the conservation of isochore patterns during this time interval. A very similar result was obtained when comparing orthologous genes from birds belonging to different avian orders (Kadi et al., 1993; Mouchiroud and Bernardi, 1993).

The correlations between GC\textsubscript{3} values of mammals (and birds) from different orders are very important because GC\textsubscript{3} values are, in turn, correlated with the GC levels of surrounding DNA segments about 200 kb in size. Since in both cases correlation coefficients are very high (0.88 in Fig. 11.2, 0.82 in the GC\textsubscript{3} vs. isochore GC correlations of human genes; see Clay et al., 1996), eq. 2 of Part 10, Chapter 2, applies. Transitivity of correlations then leads to the conclusion that long DNA stretches, at least 200 kb in size, around orthologous genes are compositionally correlated with each other in mammals and birds. In other words, the conservative mode of mammalian and avian evolution also concerns compositional patterns, as already realized many years ago (Bernardi et al., 1988). A detailed example of this situation is given in Chapter 4.2.

2.2. The conservative mode of evolution and codon usage

Another way to investigate the conservative mode of evolution is by plotting the average frequency of codons from orthologous genes. Fig. 11.4 presents the results obtained for the human/calf comparison. For reasons which will become clear in the following chapter, the gene data set was split into a GC-poor (GC\textsubscript{3} < 60\%) and a GC-rich (GC\textsubscript{3} > 60\%) subset. The two difference histograms (human values less calf values) so obtained are almost indistinguishable from the baseline. In other words, the average frequency of codons (and, therefore, also of aminoacids) from all orthologous genes that were compared is the same in the two species.

On the other hand, as we first noted (Bernardi et al., 1985b), "a different codon strategy is used for different genes located in the same genome", and such different strategy "is mainly determined by the location of genes in heavy or light isochores". This was shown by demon-

Figure 11.4. Histograms comparing average codon frequencies of orthologous genes from human and calf (Crueviller and Bernardi, unpublished results).
GC-poor genes

Homo sapiens

Bos taurus

GC-rich genes

Homo sapiens

Bos taurus
strating a correlation between GC\textsubscript{3} and the GC levels of the isochores in which the genes were embedded (see Part 3, Chapter 3). Another way to demonstrate this point is to construct the difference histogram of average codon frequencies from GC-rich and GC-poor human genes (see Fig. 11.5). A more detailed discussion of codon usage is presented in Section 3.2.

2.3. Mutational biases in the human genome

The approach used in our laboratory to infer and compare the underlying pattern of mutations among genes located in different isochores consisted in using mutation data bases that are available for many genes associated with human genetic diseases. The assumption of this approach is that such mutations reflect the actual frequency distribution of mutations as they arise in populations. It should be noted that these mutations were not yet subjected to the sieving effect of selection, in contrast to the pattern of nucleotide substitutions that depends upon two processes, mutation and fixation, the latter being influenced by selection. Inferring the mutational pattern from the observed pattern of mutations responsible of genetic alterations is not new. In fact, previous investigations
have used this kind of information for analysing the mutational dynamics of human genes (see Krawczak and Cooper, 1996).

In order to investigate whether the pattern of mutations differs among genes having different GC3 levels, we analyzed the mutational spectrum of genes with different average GC3 levels. Fig. 11.6 shows that there is very little if any variation in the expected GC content of a neutral sequence evolving only under the effect of mutations. Interestingly enough, the results show that, independently of whether CpGs, and/or repeated elements are included, the variations of the expected GC content for sequences evolving only under

Figure 11.6. The observed GC3 of each gene was plotted against the GC3 level that would be expected if it depended solely on the mutational pattern. There are four different expectations for each gene, A when repeated elements are considered, with CpG included (blue circles), or excluded (red squares); B when repeated elements are not considered, with CpG sites taken into account or not (symbols as for A). The diagonal shows the line where points should be located if the GC3 level was determined by the mutational input, so that the vertical distance between each point and this line indicates how far away is the observed GC3 from the mutational equilibrium. The horizontal lines at GC = 0.5 indicate the borders between AT bias and GC bias. (Modified from Alvarez-Valin et al., 2002).
the effect of mutations are always in the region of AT biases (with the exception of one of the results obtained from PAX6).

Fig. 11.6 also indicates that the mutational patterns in GC_{3}-rich genes do not appear to explain their GC_{3}-richness. Indeed, GC_{3}-rich and very GC_{3}-rich genes exhibit patterns of mutations that yield expectations of neutral GC level that are much lower than the actual GC_{3}. In fact, the actual GC_{3} values are between two and three times higher than the expectations obtained from mutational data. Therefore, Fig. 11.6 strongly suggests that the variability in GC_{3} cannot be attributed to variations in the mutational input along the genome but to fixational biases. In other words, even if GC_{3}-rich genes are subjected to AT biases these genes remain GC_{3}-rich, indicating that the majority of new mutations that are introduced in the populations, which are GC → AT, are eliminated. This conclusion also suggests that the compositional heterogeneity of the isochore pattern is under selective constraints since, whatever factor is responsible for maintaining high GC_{3} levels, if affects the whole isochore, as indicated by the strong correlation between the GC level of the isochore and GC_{3}.

The results of Fig. 11.6 are in agreement with recent analyses on single nucleotide polymorphisms (Eyre-Walker, 1999; Smith and Eyre-Walker, 2001). These authors found that, contrary to what would be expected if the GC content was solely determined by mutational biases (and assuming the equilibrium condition holds), the number of segregating GC → AT polymorphisms is significantly higher than that of AT → GC polymorphisms. Obviously, if the sequences were in mutational equilibrium, the two kinds of polymorphisms should be present in equal amounts.

The problem under consideration here has been investigated also through other approaches. The analysis of substitution patterns in pseudogenes can be used to infer the underlying pattern of mutations since pseudogenes are supposed to be quite free of selectional constraints. While some analyses suggest that all pseudogenes (from both GC-poor and GC-rich isochores) are under AT biases (Gojobori et al., 1982), a more recent paper on two globin pseudogenes located in different isochores (Francino and Ochman, 1999) proposed that the patterns of substitutions are not the same. According to these authors, the expected GC level that is obtained from the β globin gene (located in a GC-poor isochore) would be 0.4, whereas the expected GC level inferred from the α globin pseudogene (located in a GC-rich isochore) would be 0.57, suggesting a GC bias in GC-rich genes and an AT bias in GC-poor genes. This conclusion is contradicted by the results on the AT bias obtained on genes, pseudogenes, repeated sequences and single nucleotide polymorphisms and, therefore, the general conclusion by Francino and Ochman (1999), that isochores result from mutation and not from selection, is certainly not supported.
The two major compositional shifts in vertebrate genomes

3.1. The major shifts

Two major, intragenomic, compositional shifts took place in the genomes of the ancestors of present-day mammals and birds, respectively (see Fig. 11.1). These shifts, originally observed at the DNA level, consist in the appearance in the genomes of warm-blooded vertebrates of a relatively small percentage (10–15%) of GC-rich DNA molecules (the “genome core”) that are GC-poorer in the genomes of most cold-blooded vertebrates (Fig. 3.17).

The GC-richest 10–15% of the genomes of cold-blooded vertebrates hybridize, however, single-copy DNA from human H3 isochores (Bernardi, 1995; see Fig. 5.6). This, as well as other lines of evidence (see Part 5, Section 1.3), indicate that in cold-blooded vertebrates there is an “ancestral genome core” located in the GC-richest fractions of the genome. These GC-richest fractions are, however, not as GC-rich as the corresponding fractions of the genomes of warm-blooded vertebrates. Since the genomes of mammals and birds derive from those of two different lines of ancestral reptiles (see Section 3.3), these findings indicate that two major compositional changes independently occurred in the distinct ancestral lines leading to warm-blooded vertebrates, and that they concerned a small part of the genome, which is, interestingly, the gene-richest part of it. In warm-blooded vertebrates, the majority of the genome which did not undergo the compositional transition was called the genome desert (or the empty quarter) or the paleogenome, the minority which did, the “genome core”, or the neogenome (Bernardi, 1989; see Fig. 11.7). The paleogenome/neogenome terminology should, however, be abandoned, because it may

Figure 11.7. Scheme of the compositional transition shown by the genomes of cold- to warm-vertebrates. The “empty quarter” (or “paleogenome”) is GC-poor and gene-poor (blue box) and did not undergo any compositional change. The gene-dense, moderately GC-rich “ancestral genome core” (pink box) underwent a compositional change into a gene-dense, GC-rich “genome core” (or “neogenome”; red box).
convey the idea that the neogenome is a novel compartment of the genome instead of a compositionally modified compartment.

The two major shifts were then detected at the coding sequence level (Bernardi et al., 1985b, 1988; Perrin and Bernardi, 1987; Mouchiroud et al., 1987, 1988; Bernardi and Bernardi, 1991). Indeed, coding sequences from cold-blooded vertebrates are relatively homogeneous from a compositional viewpoint and generally characterized by low GC levels, whereas coding sequences from warm-blooded vertebrates are compositionally much more heterogeneous and reach very high GC levels (Fig. 3.17), up to 100% GC in the third codon positions of avian genes.

Another evidence for the major shifts was obtained by comparing the nucleotides in third codon positions of orthologous genes from human and *Xenopus* (as initially shown by Perrin and Bernardi, 1987; Mouchiroud et al., 1987, 1988; and Bernardi and Bernardi, 1991). When the orthologous genes of human/*Xenopus* were investigated in their GC, or in their G and C values (Fig. 11.8), points were scattered about the diagonal in the low GC

---

Figure 11.8. Correlation between G and C values of orthologous genes from human, or chicken, and *Xenopus*. The orthogonal regression lines are shown together with the diagonals (dashed lines). The equations of the regression lines and the correlation coefficients are indicated. N is the number of gene pairs explored. (From Bernardi, 2000b).

Figure 11.9. **A.** Codon frequencies of human (top) and *Xenopus* (middle) GC-poor (0-60% GC) orthologous genes. The bottom histogram shows the differences in codon frequencies. **B.** Codon frequencies of human (top) and *Xenopus* (middle) GC-rich (60-100% GC) orthologous genes. The bottom histogram shows the differences in codon frequencies. (From Cruveiller et al., 2000).
A

GC-poor genes

Homo sapiens

Xenopus laevis

B

GC-rich genes

Homo sapiens

Xenopus laevis
306

A

GC-poor genes

Gallus gallus

Xenopus laevis

B

GC-rich genes

Gallus gallus

Xenopus laevis
range, showing no directional changes between the two species, whereas human gene values were increasingly higher, on the average, compared to the corresponding Xenopus gene values, as GC3 values increased.

This produced regression lines having slopes that were close to 2. Very similar results were obtained when comparing orthologous genes from chicken and Xenopus (Fig. 11.10). Since very good correlations exist between GC3 and GC levels of extended DNA regions flanking the coding sequences (Bernardi et al., 1985b; Clay et al., 1996), the results of Fig. 11.10 also confirm that large regions surrounding orthologous genes underwent (or, alternatively, did not undergo) the compositional transition (Bernardi, 1995).

Two additional observations were reported concerning the compositional transitions under consideration. First, codon frequencies and codon usage were essentially unchanged for the orthologous genes that had not been affected by the major shift, whereas they were drastically changed for those that had been affected (Cruveiller et al., 2000). In this analysis, the sets of orthologous gene pairs from human/Xenopus and chicken/Xenopus were split into two sub-sets according to the GC3 levels of the GC-richer organism: the “GC-poor genes” (0%-60% GC3), and the “GC-rich genes” (60%-100% GC3). As far as GC3-poor genes are concerned, the results are very clear-cut in that differences in amino acids and codon usage between Xenopus and human or chicken are extremely small, confirming the data of Cruveiller et al. (1999). In other words, the average levels of amino acids and the average codon frequencies of the two sets of orthologous data are very similar and the hydrophathy of the proteins was the same (Figs. 11.9, 11.10). In contrast, GC3-rich genes of human and chicken show remarkable differences in the levels of encoded amino acids and codon frequencies relative to Xenopus and an increase in the hydrophobicity of proteins (Figs. 11.9, 11.10).

The second observation was that the regression lines between GC3, G3 and C3 values of orthologous genes from human and chicken showed a high correlation coefficient and coincided with the diagonal, indicating that the genes that had not undergone the transition and those that had, were, to a large extent, the same sets of genes in the two species (Fig. 11.11). Expectedly, the scatter about the diagonal was larger than in the corresponding calf/human plot of Fig. 11.2, but difference histograms are close to the baseline when chicken and human data are compared (Fig. 11.12).

In the last case, it is remarkable that codon choices are the same even for quartet codons. One might imagine that the similarity in codon preferences between human and chicken for each set of genes is nothing but the predictable result of the fact that genes with similar GC3 levels are compared. One might then expect for the case of GC3 data set, that G- and C-ending codons exhibit higher values in both species. However, Fig. 11.12 shows that codon frequencies in quartets is the same for both species. For instance in the leucine codon group, CTG was by far the most frequent codon in both species, while CTC was the second in the order of frequency, whereas in the glycine codon group, GGC is the most frequent codon in both species. The same comparison for the remaining codon groups and

Figure 11.10. A. Codon frequencies of chicken (top) and Xenopus (middle) GC-poor (0-60% GC3) orthologous genes. The bottom histogram shows the differences in codon frequencies. B. Codon frequencies of chicken (top) and Xenopus (middle) GC-rich (60-100% GC3) orthologous genes. The bottom histogram shows the differences in codon frequencies. (From Cruveiller et al., 2000)
also for the case of GC3-poor data set clearly indicates that the similarity in codon frequencies go much beyond what would be expected from the simpler similarity in GC3 levels.

It should be stressed that CsCl profiles and GC3 plots of avian genomes indicate an extremely high level of similarity among different orders (Kadi et al., 1993; Mouchiroud and Bernardi, 1993). This similarity also applies to the mammalian genomes belonging to the “general pattern” (Sabeur et al., 1993; Mouchiroud and Bernardi, 1993), and to the “murid pattern” (Douady et al., 2000), respectively. In other words, what has been shown here for human and chicken genes is generally valid for mammalian and avian genomes.

To summarize, (i) the compositional transitions concerned the genes and the intergenic sequences that are located in the GC-richest isochores of the genomes of warm-blooded vertebrates, namely in the isochores that correspond to the gene-dense regions of the vertebrate genome, the “genome core”; in contrast, they did not concern the genes and the intergenic sequences from the gene-poor “genome desert” (see Bernardi, 2000a), which remained compositionally stable, on the average, since the early vertebrates; (ii) the transitions occurred (and were similar) in the independent ancestral lines of mammals and birds, but in almost no cold-blooded vertebrate (the heterogeneous genomes of reptiles will be discussed later; only very few cold-blooded vertebrates have relatively GC-rich genomes, but these genomes are compositionally much more homogeneous than the genomes of warm-blooded vertebrates; see Bernardi and Bernardi, 1990a,b and Part 4); (iii) the transitions stopped before the appearance of present-day mammals and birds, as indicated by the very similar patterns found in different mammalian and avian orders, respectively (see the human/calf comparisons of Fig. 11.2 for an example).

These findings indicate that the compositional transitions affecting the “genome core” of...
the ancestors of mammals and birds, had already reached a **compositional equilibrium** at least at the times of appearance of present-day mammals and birds, and that, from those times on, the compositional patterns resulting from the cold- to warm-blooded transitions
were maintained until present (except for some limited “whole-genome” shifts; see Chapter 5). This conservation over 100 million years is remarkable, if one considers that about half of the human genes (the genes of the “genome core”) had undergone a compositional transition in which extremely high G₃ and C₃ values were reached.

3.2. Compositional constraints and codon usage

When it was realized that synonymous codons were not randomly used, the first interpretation was that codon usage was due to the optimization of codon-anticodon interaction energy (Grosjean et al., 1978). Two years later, the genome hypothesis of Grantham et al. (1980) proposed that the factors governing synonymous codon preference were not the same in different organisms and that there was a genome strategy of codon usage. Subsequently, it was shown that biased codon usage was not only species-specific, but also varied within the genome of *Escherichia coli*. This was explained as the effect of natural selection acting at the level of translation. This translational selection resulted in the preferential usage of a subset of “major codons” by highly expressed genes (Grantham et al., 1981; Gouy and Gautier, 1982), an idea that gained support from the finding that in *E. coli* and *S. cerevisiae*, the preferred codons in highly expressed sequences are recognized by the most abundant tRNAs (Ikemura 1981; 1982; 1985). A widely accepted explanation for the unequal usage of synonymous codons in microorganisms then became a balance between the forces of selection and mutation drift in a finite population (Bulmer, 1991).

Our results on codon usage in the vertebrate genome led to the discovery of compositional constraints (due, in our view, to natural selection; see Part 12). Indeed, we reported that “a different codon usage (i) is the rule for different (compositional) classes of genes from the same warm-blooded vertebrate genome; and (ii) is mainly determined by the location of genes on heavy or light isochores. This compositional constraint predominates over other constraints, which may also be operational, and be responsible for the deviations of some genes from the general relationship” (Bernardi et al., 1985b). This effect is shown in Fig. 11.5, which compares average codon usage in human GC-poor and GC-rich genes. Interestingly, this factor had been detected, but not taken into consideration, by Grantham et al. (1980).

It was not surprising then that, when applied to mammalian sequences, multivariate statistical analysis revealed a single major trend which was strongly correlated with the GC₃ content of each gene. Moreover, the first axis did not discriminate aspects of gene function such as regulation during development, tissue specificity, constitutive expression, intracellular localization of the protein product, etc. (Sharp et al., 1988, 1995; Sharp and Matassi, 1994; Duret, 2002). The conclusion that expression levels do not influence the codon usage pattern in mammals was also drawn by analyzing expressed sequence tags (ESTs) in different tissues (Duret and Mouchiroud, 2000).

Other factors do, however, play a role in shaping codon usage, even in the case of vertebrates. Indeed, if multivariate analysis is applied to the coding sequences of *Xenopus*, which are characterized by a rather average GC₃ distribution, with very scarce high values, the results obtained (Musto et al., 2001) show that codon usage, although mainly determined by compositional constraints, is, in fact, also influenced by translational selection
(Fig. 11.13). In fact, a second trend, independent of GC$_3$ discriminates highly expressed sequences (mainly housekeeping, but also tissue-specific) from the other sequences. This is confirmed by the significant correlation found between the position of each gene along the second axis, and the numbers of matching ESTs.

Our early results on vertebrates led us to consider compositional constraints in microorganisms (and viruses) as another factor determining codon usage next to translational selection. The correlation of GC$_3$ with genome GC exhibited by prokaryotes, showed that, indeed, compositional constraints also played a role in this case (Bernardi and Bernardi, 1985, 1986a; Muto and Osawa, 1987).

One can explain the initial emphasis on translational selection as due to the fact that the first sets of available genes were from the genomes of *E. coli* and *S. cerevisiae*, which happen to be characterized by 50% and 40% GC levels, respectively. In this case, compositional constraints are low, and codon usage in microorganisms was, therefore, visualized as mainly determined by translational selection. Later on, codon usage was rather seen as due to an equilibrium between translational selection and mutational bias (considered to be responsible for base composition), an idea that was later extrapolated from microorganisms to *Drosophila melanogaster* (Shields et al., 1988; Sharp and Li, 1989; Moriyama and Gojobori, 1992; Carulli et al., 1993; Akashi, 1994, 1997; Moriyama and Powell, 1997; Powell and Moriyama, 1997), *Caenorhabditis elegans* (Stenico et al., 1994), *Zea mays* (Fennoy and Bailey-Serres, 1993) and *Arabidopsis thaliana* (Chiapello et al., 1998).

In our view, genome composition, both in prokaryotes and eukaryotes, is the result of natural selection (see Part 12). This main selective constraint may be rather uniform, as is the case for most prokaryotic genomes and for compositionally homogeneous eukaryotic genomes (like the yeast genome), or may vary within a genome, when the genome is compositionally heterogeneous. The compositional effect can be so powerful as to make very difficult to detect weaker selective forces, (e.g., translational selection). In turn, this explains why genomes that are extremely biased in terms of composition can display very similar global codon usages even when they are separated by enormous phylogenetic distances, as it happens, for example, for *Plasmodium falciparum* and *Staphylococcus*.
TABLE 11.1
Relative Synonymous Codon Usage (RSCU) values of *P. falciparum* (*Pf*) and *S. aureus* (*Sa*)

<table>
<thead>
<tr>
<th>aa</th>
<th>Codon</th>
<th>Pf</th>
<th>Sa</th>
<th>aa</th>
<th>Codon</th>
<th>Pf</th>
<th>Sa</th>
<th>aa</th>
<th>Codon</th>
<th>Pf</th>
<th>Sa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phe</td>
<td>TTT</td>
<td>1.6</td>
<td>1.4</td>
<td>Ser</td>
<td>TCT</td>
<td>1.4</td>
<td>1.4</td>
<td>Tyr</td>
<td>TAT</td>
<td>1.8</td>
<td>1.6</td>
</tr>
<tr>
<td>Phe</td>
<td>TTC</td>
<td>0.4</td>
<td>0.6</td>
<td>Ser</td>
<td>TCC</td>
<td>0.4</td>
<td>0.2</td>
<td>Tyr</td>
<td>TAC</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>Leu</td>
<td>TTA</td>
<td>4.1</td>
<td>3.3</td>
<td>Ser</td>
<td>TCA</td>
<td>1.9</td>
<td>1.8</td>
<td>End</td>
<td>TAA</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Leu</td>
<td>TTG</td>
<td>0.7</td>
<td>0.9</td>
<td>Ser</td>
<td>TCG</td>
<td>0.2</td>
<td>0.3</td>
<td>End</td>
<td>TAG</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Leu</td>
<td>CTT</td>
<td>0.7</td>
<td>0.8</td>
<td>Pro</td>
<td>CCT</td>
<td>1.3</td>
<td>1.5</td>
<td>His</td>
<td>CAT</td>
<td>1.6</td>
<td>1.5</td>
</tr>
<tr>
<td>Leu</td>
<td>CTC</td>
<td>0.1</td>
<td>0.2</td>
<td>Pro</td>
<td>CCC</td>
<td>0.3</td>
<td>0.2</td>
<td>His</td>
<td>CAC</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>Leu</td>
<td>CTA</td>
<td>0.3</td>
<td>0.6</td>
<td>Pro</td>
<td>CCA</td>
<td>2.4</td>
<td>1.8</td>
<td>Gln</td>
<td>CAA</td>
<td>1.8</td>
<td>1.7</td>
</tr>
<tr>
<td>Leu</td>
<td>CTG</td>
<td>0.1</td>
<td>0.2</td>
<td>Pro</td>
<td>CCG</td>
<td>0.1</td>
<td>0.5</td>
<td>Gln</td>
<td>CAG</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>Ile</td>
<td>ATT</td>
<td>1.4</td>
<td>1.6</td>
<td>Thr</td>
<td>ACT</td>
<td>1.3</td>
<td>1.2</td>
<td>Asn</td>
<td>AAT</td>
<td>1.7</td>
<td>1.4</td>
</tr>
<tr>
<td>Ile</td>
<td>ATC</td>
<td>0.2</td>
<td>0.5</td>
<td>Thr</td>
<td>ACC</td>
<td>0.5</td>
<td>0.2</td>
<td>Asn</td>
<td>AAC</td>
<td>0.3</td>
<td>0.6</td>
</tr>
<tr>
<td>Ile</td>
<td>ATA</td>
<td>1.4</td>
<td>0.9</td>
<td>Thr</td>
<td>ACA</td>
<td>2.0</td>
<td>2.1</td>
<td>Lys</td>
<td>AAA</td>
<td>1.7</td>
<td>1.6</td>
</tr>
<tr>
<td>Met</td>
<td>ATG</td>
<td>1.0</td>
<td>1.0</td>
<td>Thr</td>
<td>ACG</td>
<td>0.2</td>
<td>0.5</td>
<td>Lys</td>
<td>AAG</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Val</td>
<td>GTT</td>
<td>1.8</td>
<td>1.6</td>
<td>Ala</td>
<td>GCT</td>
<td>1.9</td>
<td>1.4</td>
<td>Asp</td>
<td>GAT</td>
<td>1.8</td>
<td>1.6</td>
</tr>
<tr>
<td>Val</td>
<td>GTC</td>
<td>0.2</td>
<td>0.4</td>
<td>Ala</td>
<td>GCC</td>
<td>0.4</td>
<td>0.3</td>
<td>Asp</td>
<td>GAC</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>Val</td>
<td>GTA</td>
<td>1.7</td>
<td>1.5</td>
<td>Ala</td>
<td>GCA</td>
<td>1.6</td>
<td>1.9</td>
<td>Glu</td>
<td>GAA</td>
<td>1.8</td>
<td>1.6</td>
</tr>
<tr>
<td>Val</td>
<td>GTG</td>
<td>0.3</td>
<td>0.5</td>
<td>Ala</td>
<td>GCG</td>
<td>0.1</td>
<td>0.4</td>
<td>Glu</td>
<td>GAG</td>
<td>0.2</td>
<td>0.4</td>
</tr>
</tbody>
</table>

* From Musto et al. (1995). *Pf* and *Sa* are *P. falciparum* and *S. aureus*, respectively. RSCU were calculated according to Sharp et al., 1986. Asterisks indicate stop codons; these were not taken into account.

*aureus* (see *Part 9* and Table 11.1), although some translational selection is still detectable (Musto et al., 1999).

To sum up, codon usage is due to two main selective factors: 1) compositional constraints (which we visualize as the result of natural selection shaping up the compositional patterns of genomes), and 2) translational selection. When the former are very biased, the latter becomes barely detectable (one should remember that at 100% GC, 50% of codons are not used) and, vice versa, when the compositional constraints are not biased (and genomes are in middle GC range), translational selection becomes quite visible.

The importance of compositional constraints on codon usage may be demonstrated in another way. As shown in Figs. 11.9A and 11.10A, there is no difference in average codon frequency between orthologous GC-poor genes of *Xenopus* and human or chicken. In other words, the average codon frequency has been conserved in GC-poor genes since at least the common ancestor of tetrapods. (In fact, data by Romero et al., 2003, indicate a conservation dating back to Cyprinids, a fish family belonging to the superorder *Ostariophysi*; see *Part 4*). In sharp contrast, Figs. 11.9B and 11.10B show large differences in average codon frequency in GC-rich genes of *Xenopus* and human/chicken. Again, however, average codon frequencies of GC-rich genes are very close to each other in warm-blooded vertebrates (human and chicken; Fig. 11.12B).
3.3. Other changes accompanying the major shifts

The compositional transitions that took place in the genomes of the ancestors of mammals and birds involved more than the compositional changes just described. Indeed, (i) **CpG islands** were formed (See Part 5, Chapter 2); these regulatory sequences, about 1 kb in size, mainly located 5' of GC-rich genes and characterized by high levels of GC and unmethylated CpG doublets, do not exist in fishes and amphibians; interestingly they do not exist even in reptiles, like crocodile and turtle, that exhibit some compositional heterogeneity (Aissani and Bernardi, 1991a); these results rule out the proposal (Bird, 1987) that unmethylated CpG islands are remnants of the less methylated genomes of invertebrates ancestral to vertebrates; (ii) **nucleotide context changes** towards G and C took place around the AUG initiator codon of GC-rich human genes relative to GC-poor human genes and genes from cold-blooded vertebrates (Pesole et al., 1999; see Part 5. 1.2); like (i) above, this point indicates that the major shifts also caused **compositional changes in regulatory sequences**; (iii) **H3+ bands** appeared in metaphase chromosomes (see Part 7) as a consequence of the GC increase corresponding to the compositional transition; (iv) **karyotype changes** and **speciation** increased (Bernardi, 1993b), probably in association with a higher level of recombination in distal parts of chromosomes of warm-blooded vertebrates compared to cold-blooded vertebrates; it has been argued (Bernardi, 1993b) that the increased genome instability associated with the higher recombinogenicity is a disadvantage, which is compensated by a higher genetic variability, an advantage in terms of coping with environmental changes and/or exploration of new ecological niches; a similar reasoning was already made for the mitochondrial genome of yeast (see Part 2, Chapter 3.5). (v) **DNA methylation** and **CpG doublet concentration** decreased by a factor of two in mammals and birds, compared to fishes and amphibians (Jabbari et al., 1997; see Part 5, Chapter 3, and the following section).

This methylation/CpG transition deserves a more detailed discussion. An analysis of the data presented in Part 5 indicates that (i) two positive correlations hold between the 5mC and GC levels of the genome of fishes/amphibians and mammals/birds, respectively (see Fig. 5.13); (ii) the higher methylation of fishes and amphibians is not related to the higher amounts of repetitive DNA sequences (see Fig. 5.14); and (iii) the 5mC and CpG observed/expected values show no overlap between the two groups of vertebrates and suggest the existence of two **equilibria in 5mC and CpG levels** (see Fig. 11.14). Several important questions then arise concerning (i) the two equilibria in methylation and CpG shortage; (ii) the transition between the two equilibria; and (iii) the causes of the methylation/CpG transition.

The **existence of two equilibria** of DNA methylation and CpG in vertebrate genomes is stressed by the very similar methylation levels of mammals belonging to orders separated from each other by 100 million years. As already mentioned in Part 5, because of the star-like phylogeny of mammalian orders (see Fig. 11.3), this situation indicates that the ancestral mammalian genome also showed low CpG (o/e) and 5mC levels. This conclusion contradicts the hypothesis (Cooper and Krawczak, 1989) that the vertebrate genome originally was strongly methylated (and GC-rich) and that the methylated CpG doublets (as well as the GC levels) subsequently underwent a monotonous decay (see Fig. 11.15). An additional difficulty of that model was that the ancestors of vertebrates, the cephalochor-
Figure 11.14 Plot of CpG observed/expected ratio (CpG o/e) against 5mC level in the genomes of cold-blooded and warm-blooded vertebrates. CpG o/e values were taken from the available literature for four cold-blooded vertebrates (three fishes and one amphibian) and nine warm-blooded vertebrates (eight mammals and one bird). All available methylation data were used, namely 32 cold-blooded vertebrate and 46 warm-blooded vertebrate species, respectively. Horizontal and vertical lines crossing the boxes correspond to standard deviations. Note the reverse scale on the 5mC axis. This was done to put fishes/amphibians and mammals/birds in an essentially time perspective. (From Jabbari et al., 1997).

Figure 11.15. The results of Jabbari et al. (1997) on DNA methylation of fishes/amphibians (upper horizontal line) and of reptiles/mammals/birds (lower horizontal line) and the average GC level (horizontal line in the bottom frame) are compared with the monotonous decrease of CpG (red line) and GC (blue line) hypothesized by Cooper and Krawczak (1989). (Modified from Jabbari et al., 1997).
dates and urochordates, are characterized by genomes that are neither GC-rich nor strongly methylated, making it difficult to understand how and when the hypothesized original high GC and methylation of vertebrates had originated. The considerations just presented also contradict the hypothesis of an equilibrium between CpG depletion and CpG formation (Sved and Bird, 1990) which also predicts a strong decrease in CpG during vertebrate evolution (this decrease being, however, less rapid than that predicted by Coop-
er and Krawczak, 1989).

Concerning the **methylation/CpG levels changes**, the results available so far suggest that reptiles are characterized by a lower methylation level, compared to fishes and amphibians, close to those of mammals and birds placing the methylation transition at the appearance of the common ancestor of reptiles and mammals, namely at the appearance of amniotes.

One should, however, keep in mind that the number of reptilian DNAs analysed so far is small, that rather different levels were found in the few reptiles investigated and that no estimate is available for the CpG shortage. Under these circumstances, the question as to whether the low methylation level is present in all or only in some reptiles is still open. Therefore, two possibilities should be considered, namely, that either the common reptilian ancestor of mammals and birds had already undergone a decrease in methylation which was then transmitted by descent to extant reptiles, mammals and birds; or that the methylation change occurred independently in reptiles (possibly, only in some of them), in mammals and in birds.

As far as the **causes** for the methylation/CpG transition are concerned, they will be discussed in **Part 12**, along with the causes for the major shifts.
CHAPTER 4

The minor shift of murids

4.1. Differences in the compositional patterns of murids and other mammals

Although characteristic differences between the CsCl profiles of murids and other mammals were first observed by Thiery et al. (1976; see Table 3.2) only further more detailed investigations definitely proved that the mouse genome did not exhibit the GC-richest DNA components that were present in the human genome (Salinas et al., 1986; Zerial et al., 1986a). These findings were confirmed by investigations at the coding sequence level, which demonstrated in murid genes characterized by extreme base compositions (highest and lowest GC levels) a minor, intragenomic, shift (see Fig. 11.16) relative to orthologous genes from mammals exhibiting the general pattern (Mouchiroud et al., 1988). Indeed, the GC-richest and GC-poorest genes of the murids are less GC-rich and less GC-poor, respectively, than their orthologs from other mammals, whereas the genes with intermediate GC values are unchanged. The very high correlation coefficient of the human/mouse plot of Fig. 11.16 indicates a conservation of the order of GC levels in orthologous genes from human and murids (Mouchiroud et al., 1988). Now, it is known that murids exhibit (i) rates of synonymous substitutions that are higher, by a factor of 5 to 10, relative to human coding sequences (Wu and Li, 1985; Gu and Li, 1992); (ii) a defective repair system (see Holliday, 1995); and (iii) a compositional pattern that is derived from the general mammalian pattern (Galtier and Mouchiroud, 1998).

As far as the last point is concerned, if murids arose relatively late in mammalian

![Figure 11.16](image-url)

**Figure 11.16.** Correlations between GC₃ values of orthologous genes from human and calf and human and mouse. The orthogonal regression lines practically coincide with the diagonal (broken line) in the case of calf, but diverge in the case of mouse. The equations of the regression lines and the correlation coefficients are shown. (From Bernardi, 2000a).
evolution, as indicated by the paleontological record (Carroll, 1988), it was difficult to escape the conclusion that their genome underwent a narrowing of its compositional distribution compared to the mammals from which they were derived (Mouchiroud et al., 1988). Molecular data were interpreted, however, to suggest (Easteal, 1990; Li et al., 1990; Bulmer et al., 1991) that Myomorpha branched off before the divergence (at 100 Myrs) among carnivores, lagomorphs, artiodactyls and primates, a divergence of mammalian orders which all share the general pattern. The absence of a fossil record for over 50 Myrs obviously outweigh the molecular data and favour a derived compared to a primitive murid pattern (Sabeur et al., 1993).

Under these circumstances, it is interesting to observe that the higher mutational input "randomizes" the composition of synonymous positions by reducing the difference between the extreme GC values found in the general mammalian pattern. Indeed, under a mutational bias model à la Francino and Ochman (1999), one would expect instead an increase of the differences between extreme values. It is also remarkable that this "randomization" process also concerns non-coding sequences, and even regulatory sequences, causing for instance, an "erosion" of CpG islands (Aissani and Bernardi, 1991a,b; Matsuo et al., 1993). In the case of α globin, for example, the human gene is very GC-rich and is associated with a CpG island, whereas the mouse gene is less GC-rich and has lost the island.

4.2. Isochore conservation in MHC loci of human and mouse

An important question concerns the effect of the minor shift on the isochore pattern. This problem has been explored on all syntenic regions shared by human and mouse, that were large enough to comprise more than one isochore. The results of this investigation (Pavliček et al., 2002b) indicate that the human isochore pattern is still recognizable in the mouse chromosomes after the minor shift. This will be illustrated here by considering the major histocompatibility (MHC) locus.

The MHC region of mammalian genomes embeds a series of highly polymorphic genes that are expressed on the surface of several cell types, including the T and B lymphocytes that form an integral part of the adaptive immune system. In human, the MHC (or HLA) locus covers roughly 4 Mb of chromosome 6, while in mouse the MHC (or H2) locus is located in a largely syntenic region on chromosome 17. The locus has long been of intrinsic interest in view of its central role in conferring immunity, of its polymorphism, and of its paralogous loci on human chromosomes 1, 9 and 19 (Endo et al., 1997; MHC Sequencing Consortium, 1999). More recently, it has provided a paradigmatic example of the sharp boundaries that can exist between mammalian isochores. The MHC region in human has also provided the first examples of a precise link between isochore boundaries and replication timing switchpoints at the sequence level. The central, GC-poor isochore (Fig. 11.17, blue and yellow landscape for human) is replicated late, the two GC-rich isochores that flank it (orange and red) are replicated early, and the sharp boundaries of the central isochore both coincide with the switches in replication timing (Tenzen et al., 1997; the MHC Sequencing Consortium, 1999; and references therein). This region therefore provides a particularly clear illustration of the genome-wide tendency of GC-rich regions (and
Figure 11.17. Color-coded GC scans, sequence similarity dot plot and repetitive element contents for homologous parts of human and mouse MHC loci. In both species the scale increases from the proximal (centromeric, upper/left) beginning of the sequence to its distal (telomeric) end. GC levels were plotted using overlapping 100 kb sliding windows and partitioned into five color-coded intervals, representing the five isochore families in human: their limits, indicated by horizontal dashed lines, are as in Bernardi (2001). Bright colors correspond to dot plot diagonals and show extended sequence homology, dull colors show regions with no or only local homology. In both species, the locus spans three distinct isochores, characterized in human by red and orange for the two GC-rich isochores, and by blue and yellow for the central GC-poor isochore. The MHC regions are indicated by bars below the GC scans. These regions correspond to the classes of MHC genes (MHC Sequencing Consortium, 1999; Stephens et al., 1999), and are blockwise syntenic between human (HLA) and mouse (H2), with the exception of the H2-K subregion in mouse (short yellow bar near the proximal end). Opposite each GC scan are the corresponding contributions of repetitive elements, which closely follow expectations from genome-wide analyses: e.g. in human the LINE1 elements are more frequent in L1 and L2 DNA, while Alus are most frequent in H2 DNA (Bernardi, 2001, and references therein). (From Pavlicek et al., 2002b).

the bands in which they are located) to replicate earlier than GC-poor regions (Federico et al., 2000).

The recent sequencing of the mouse MHC locus offers a first opportunity to compare a
contiguously sequenced, nearly syntenic region spanning three clearly delimited isochores in human and mouse. Such a comparison is of particular interest because, as already mentioned, human and mouse represent the two principal patterns of base compositional organization in eutherians: mouse and rat belong to a subset of rodents (coinciding essentially with the myomorphs) that differ from other eutherians in having less dramatic contrasts in GC level among and within their isochores (reviewed in Bernardi, 2000; see Part 11). This phenomenon also affects the compositionally prominent, typically unmethylated CpG islands, which surround the promoters of nearly all housekeeping and many tissue-specific genes: in mouse they are lower, shorter and less abundant than in human (see Part 5, Chapter 2).

As can be seen in Fig. 11.17, the isochore structure of the locus is conserved between human and mouse, not only in the regions showing sequence homology (as estimated by standard similarity criteria), but also in the intervening regions (represented by dull colors and the absence of diagonals in the dot plot). The intervening regions between the bright homology stripes include the H2-K class I sub-region in mouse, which is located between the two most proximal homology regions shown in Fig. 11.17 (~0.25 Mb from the proximal beginning of the sequence), and which is thought to have been transposed from a distant class I region near the distal end of the sequence. The H2-K insertion locally disrupts the synteny of genes, the otherwise strict partitioning of the classes into different regions and the sequence similarity between human and mouse, but it does not interrupt the homogeneity of the isochore or the compositional synteny of the locus.

The conservation of the number and extent of the isochores is not altogether trivial, since human and mouse span the largest differences in compositional genome organization that exist among eutherians (Bernardi, 2000a). Such a conservation could, however, be expected from the GC3 levels of the genes in this locus, from the strong compositional correlation that exists between the GC3 of genes and the GC levels of the much longer regions of DNA that embed them, and from the correlation between genic GC3 levels of orthologous genes in human and mouse. Furthermore, it is of interest that the general correlation between the GC levels of corresponding regions in human and mouse loci remains clearly visible (from the similar landscapes of the GC scans) even where local sequence similarity is weak or insignificant.

The GC-rich proximal parts (left for human, upper for mouse) correspond to the extended class II MHC region, the central GC-poor parts in both species correspond to the ‘classical’ (immunological) class II region, and the distal parts represent the class III region and the beginning of the class I region in human (Stephens et al., 1999). Although class III and extended class II are both GC-rich, they do not have the same degree of polymorphism: the polymorphism of the extended class II in human, although not as high as that of the classical class II, is closer to it than to the low polymorphism in class III. In line with the observations on polymorphism in human, homology between human and mouse is again typically well conserved in MHC class III, while the level of conservation is lower for the extended and especially for the very GC-poor classical MHC class II, where homology is limited to a few short regions. The MHC locus thus shows on the whole a higher level of evolutionary conservation of the GC-rich, gene-rich regions, and less conservation of the long GC-poor segment. Moreover, the lengths of the GC-rich class III and extended class II regions are similar for both human and mouse, while the GC-poor classical class II
region is significantly shorter in mouse than in human. This difference, which corresponds to a loss of GC-poor DNA in an ancestor of mouse, is in agreement with a general propensity of more compact vertebrate genomes to preferential DNA contraction in the GC-poorer, and typically gene-poorer, regions (Bernardi, 1995; see Fig. 5.14C).

The \textit{general reduction of compositional contrast in mouse} is visible also in the MHC region: the difference in GC level between the central and flanking isochores is much less pronounced than in human. The precise localization of isochore boundaries in mouse, and the prediction of corresponding replication timing switchpoints, could thus profit from comparisons with orthologous human sequences. Indeed, combined GC scan/dot plot alignments such as the one shown in Fig. 11.17 could help reveal isochore boundaries: for example, the proximal boundary of the GC-poor isochore in mouse is compositionally less sharp than the distal boundary in mouse, or than the two boundaries in human. In the case of this central MHC isochore in mouse, all comparisons (homology of sequences, homology of corresponding syntenic genes, corresponding changes in GC level) consistently point to the same location for its proximal boundary, namely within a short region (\(
\ll 100 \text{ kb}\)) delimited by two homology stripes, located approximately 350 kb into the mouse sequence, and indicated by a division in Fig. 11.17. More detailed analyses by experiments (cf. Tenzen et al., 1997) and segmentation algorithms (cf. Oliver et al., 2001) should help to confirm its precise position.

The repeat content of the human MHC sequence in Fig. 11.17 is 45.5%, Alus comprising 16.1% and LINE1 elements 15.6% of the sequence. The mouse sequence contains 6% B1 sequences (the most recent mouse SINE class, analogous to Alus in human) and 3.1% LINE1 elements. These numbers are very close to previous genome-wide estimates, at corresponding GC levels, of repeat densities in mouse and human. It is interesting that, in spite of independent insertions of these classes of retroelements, their isochore distribution is similar in both species. Human and mouse LINE1 elements are, as expected (see Part 6), more dense in the GC-poor part, namely classical class II, while SINE repeats, human Alus and mouse B1s, are more frequent in the GC-rich parts of the locus (extended class II and class III). We have proposed that negative selection on compositionally non-matching elements is the main factor influencing the distribution of LINES and SINES (see Part 6).

The functional meaning of the strict regional partitioning of MHC gene classes into different isochores, and of the conservation of this partitioning despite high recombination during the evolution of the locus, speak for a strong selection that maintains the precise isochore structure of this region.

Recent observation showed the isochore conservation between all human, mouse and feline classical and extended class II MHC regions (Yuhki et al., 2003). Despite many rearrangements and extensive gain and loss of genes during long separate evolution, the human and cat classical and extended class II regions show nearly identical GC content. The cat extended and classical class II regions maintain 49.1% and 41.2% GC content, whereas the same human regions have 49.9% and 41.3% of GC (Yuhki et al., 2003). Given the weak similarity in the majority of DNA sequences without coding capacity, the selection on the global structure and composition of large genomic regions rather than conservation of individual bases or genes, seems to be the main factor retaining the similar GC content in large genomic segments of mammals.
4.3. The increased mutational input in murids

As already mentioned, the very similar compositional patterns of genomes from mammalian orders that were separated for 100 million years indicates that the “general” mammalian pattern was already present in the common ancestor of present-day mammals and was the result of an equilibrium between the mutational input and negative selection. When the mutational input increased, as in the case of murids, a new equilibrium was reached, as witnessed by the very similar compositional pattern of different murids (Douady et al., 2000a,b; see Part 4, Chapter 5), and a conservative mode of evolution took over. This new equilibrium already existed in the common ancestor of murids and was, therefore, reached 30 million years ago or earlier. This is the approximate time of divergence for rodents that have a defective repair system and higher rates of nucleotide substitutions (like murids) compared to other rodents (like Histricomorphs and Sciuromorphs) that have a stringent repair system and correspondingly share the general mammalian pattern.
Whole-genome (or horizontal) shifts were observed in the genomes of cold-blooded vertebrates, mainly of fishes (Bernardi and Bernardi, 1990a,b). They consist in shifts of the entire distribution of DNA molecules towards higher or lower GC levels (see Fig. 11.1C). Three possible, not mutually exclusive, explanations for such shifts are (i) changes in the amounts of cryptic satellite DNAs; (ii) changes in the amounts of interspersed repeats; (iii) "mutator mutations", that is changes due to mutations in the sequences coding for protein sub-units of the replication machinery; in bacteria, these mutations lead not only to mutational biases, but also to changes in mutation rates.

(i) The first explanation can be ruled out, at least as a general explanation, since the shift is also seen in the long non-satellite sequences which are now available for some fish species (see Fig. 4.5A; Jabbari and Bernardi, 2004a).

(ii) The second explanation can also be ruled out, again at least as a general explanation, because compositional changes, in the same direction as those shown by DNA molecules, are exhibited by the coding sequences from the same fish species (Fig. 4.6A,B; Jabbari and Bernardi, 2004a).

(iii) The third explanation should, therefore, be accepted as the most general explanation for the "whole genome shift". In fact, the very existence of "mutator mutations" cannot be denied in vertebrates, because of the well-demonstrated AT bias in the human genome (see Chapter 2, Section 7). This basic explanation does not rule out some contribution from cryptic satellites and interspersed repeats to the whole genome shift of vertebrates nor possible overlaps with major intragenomic shifts (see Fig. 11.1A).

Interestingly, among vertebrates, the whole-genome shifts are more frequent and larger in fishes than in mammals and birds (Bernardi and Bernardi, 1990a, b), a difference that will be discussed in Part 12. In the case of fish genomes, different orders are generally characterized by modal buoyant densities that are different in their average values as well as in their ranges. In contrast, different families within any given order are more often characterized by narrow ranges of modal buoyant densities and no difference in modal buoyant density is found within any single genus (except for the genus Aphyosemion, which should be split into several genera).

The compositional differences that were found among species belonging to different orders and to different families within the same order are indicative of whole genome, or horizontal, compositional transitions. These transitions were found to be independent of geological time (see Fig. 11.18). Moreover, the rates of directional base substitutions responsible for the transitions were found to be very variable and to reach, in some cases, extremely high values (Bernardi and Bernardi, 1990b).
Figure 11.18. Standard deviations of average modal buoyant densitites in CsCl, $\sigma$, of DNAs from fish species belonging to the same order, family, or genus (in the only case of *Aphyosemion*) are plotted against the time of the appearance of these groups. Groups are: Lamniformes, 1; Cypriniformes, 7; Salmonidae, 8; Gadiformes, 12A; Aplocheilidae, 12B; Aphyosemion, 14; Perciformes, 14; Tetraodontiformes, 15. Numbers correspond to those of Table 2 of the original paper. (From Bernardi and Bernardi, 1990b).
Recapitulation

* In this section, some key figures from the main text (numbers in bold type with an R prefix) are represented on a small scale, with an abbreviated legend (carrying the original figure number) and no reference. See the main text figures for further information. Some other figures and tables from the main text are only referred to (they are in plain type)
1. Structural genomics of warm-blooded vertebrates

Our main conclusions on the structural genomics of vertebrates can be summarized as follows (see Table 12.1).

(i) The genome compartmentalization, i.e., a genome’s discontinuous compositional heterogeneity, is best illustrated by describing the organization of the genomes of warm-blooded vertebrates. We can take as an example the human genome, since we have shown that it is typical for most mammalian genomes and similar to the avian genomes. The first property that we discovered was its compositional compartmentalization (Fig. R1). When human DNA is randomly broken down during preparation to an average size of 100 kb, the fragments were shown to cover a broad compositional spectrum (30-60% GC) and to be distributed in a small number of families (the “major DNA components”), each one of which is comparable in heterogeneity to the least heterogeneous genomes, those of prokaryotes (Figs. 3.10, 3.11). These families, L1, L2, H1, H2, H3, represent about 33%, 30%, 24%, 7.5%, 5% of the genome. We could also show that the 100-kb DNA fragments derived from much longer, fairly homogeneous chromosomal segments (at least 300 kb in average size), which we called isochores for (compositionally) equal landscapes. Incidentally, satellite and ribosomal DNAs could also be visualized as isochores, because of their compositional homogeneity. To sum up, the genomes of warm-blooded vertebrates can be described as mosaics of isochores (Fig. R1A).

![Diagram of isochores and DNA fragmentation](image)

Figure R1 (3.9). A. Scheme of the isochore organization of the human genome. B. The CsCl profile of human DNA (ca. 100 kb) is resolved into its major DNA components.
(ii) Isochore (or compositional) patterns of vertebrate genomes were called **genome phenotypes**, because they show phylogenetic differences. The major difference was observed between the genomes of warm- and cold-blooded vertebrates, the latter consisting of isochores that cover a much narrower compositional range and attain lower GC levels. Lesser yet significant differences were found between the isochore patterns of birds (which comprise an additional very small, very GC-rich isochore family H4), or murids (which do not exhibit a distinct H3 isochore family), and the general mammalian pattern presented by the human genome (see **Fig. R2A**). The compositional distributions of coding sequences (or of their third codon positions, **Fig. R2B**) from vertebrates belonging to different classes mimic their isochore patterns in that, compared to human, *Xenopus* shows a narrower spectrum centered on low values, whereas chicken and mouse extend more and less, respectively, towards high values.

![Figure R2 (3.17). A. Isochore families as deduced from density gradient centrifugation. B. Compositional patterns of third codon positions (GC3).](image)

(iii) The genomic code comprises three kinds of relations: 1) the **compositional correlations between coding and noncoding sequences** which were found for all vertebrate genomes explored (see **Figs. R3A and 3.18, 3.19**). This was a striking result since coding sequences only represent a minute amount of those genomes (2-3% in the case of the human genome), but also provided the strongest support for the genome as a **coordinated ensemble**. 2) The functional compatibility or incompatibility between coding and noncoding sequences indicated the existence of a **genomic fitness**. This can be exemplified by the effect of compositionally matching vs. non-matching contiguous coding sequences on the transcription of integrated viral sequences (**Fig. R4**). 3) A **universal correlation** (**Fig. R5**) was found among the compositions of the three codon positions, both within and among genomes, ranging from bacteria to human.

(iv) The **distribution of genes** in the human genome is remarkably non-uniform, about
Figure R3 (3.20). A. Correlations between GC$_3$ of human genes and the GC level of DNA fractions or YACs in which the genes were localized (filled circles), or of 3' flanking sequences (open circles); B. Correlations between GC levels of human coding sequences (CDS) and of the corresponding introns.

Figure R4 (6.4). Distribution of integrated HTLV-1 and HBV viral sequences in the human genome. The CsCl profile of human DNA is shown. Arrows indicate the GC level of viral sequences. Expressed sequences are in red, non-expressed sequences in yellow.
50% of the genes being located in H2 and H3 isochores, which only represent about 12% of the genome, the other 50% or so being present in the remaining 88% of the genome formed by the GC-poorer L1, L2 and H1 isochores (Fig. R6). The gene distribution defines, in fact, two gene spaces, the small genome core and the large genome desert (or empty quarter, from the classical name of the Arabian desert).

(v) The genome core and the genome desert exhibit distinct structural and functional properties (Table 5.1; see also Section 2 below). The genome core is characterized not only by a high gene density, but also by abundant CpG islands, short introns and short untranslated sequences, early replication, high recombination, high transcription, open chromatin structures and distal chromosomal locations, whereas the genome desert is characterized by opposite features. Moreover, the open chromatin of the genome core is preferentially targeted by transposed duplicated genes and by viral sequences, which are, however, stable and transcriptionally active only in isopycnic isochores (i.e., isochores having a GC level

Figure R5 (from Fig. 10.9). Correlations between GC₃ and GC₁ or GC₂ for 20,148 human genes.

Figure R6 (5.1). A. Profile of relative gene concentration (red dots) in the human genome. B. Density of genes in isochore families.
approximating that of the proviral sequence under consideration; see Fig. R5 and Part 6). Likewise, GC-rich SINE sequences and GC-poor LINEs have their highest densities and stabilities in isopycnic isochores.

2. Chromosomes and interphase nuclei

**Compositional maps** of human chromosomes (Figs. R7 and 3.13-3.16) allow visualizing of the isochores and of the gene densities in them. As far as the distribution of isochores in prometaphase chromosomes is concerned, four subsets of bands could be distinguished (Table 7.1 and Fig. R8): two subsets, H3\(^+\) and H3\(^-\), of Reverse bands and two subsets, L1\(^+\) and L1\(^-\), of Giemsa bands. H3\(^+\) bands are the R bands that hybridize H3 and H2 isochores; they largely correspond to the T bands of Dutrillaux (1973) and are preferentially located at telomeric and distal regions of chromosomes. L1\(^+\) bands are the G bands that hybridize L1 isochores; they correspond to the two sets of the most intense staining G bands of Francke (1994) and tend to be located in the central parts of chromosomes. GC levels decrease from H3\(^+\) to H3\(^-\) R bands and then to L1\(^-\) and L1\(^+\) G bands, in this order. H3\(^-\) and L1\(^-\) bands are, however, not very different compositionally, (in fact, H3\(^-\) bands exist that are GC-poorer than non contiguous L1\(^-\) bands), but G bands are always flanked by GC-richer R bands, and R bands by GC-poorer G bands in chromosomes 21 and 22. Moreover, DNA is more tightly packed in G bands than in R bands.

In **interphase nuclei**, H3 isochores are in a much more expanded chromatin configuration compared to the compact chromatin of L1 isochores (Figs. R9, 7.27). Moreover, H3 isochores are located in the center of the nucleus, whereas L1 isochores are packed against the nuclear membrane.

---

Figure R7 (7 13A) Bands, isochores and gene concentration of human chromosomes 21 and 22. Bands comprise four classes of G bands characterized by different staining intensity (from black to pale grey) and two classes of R bands (H3\(^+\), red; H3\(^-\), white); the two chromosomes are represented according to their cytogenetic size. GC profiles are shown through 100 kb windows. Genes/Mb, the blue bar plot concerns chromosomal bands, the red plot 1-Mb segments.
Figure R8 (7.12). Identification of the GC-poorest (L1+, blue bars) and the GC richest (H3+, red bars) human chromosomal bands. The H3 R bands are in white. The grey scale of the L1 G bands is according to Francke (1994).

Figure R9 (7.27). A cartoon showing that the centromere-proximal regions of chromosomes tend to be present in compact chromatin structures at the periphery of the interphase nucleus (blue blocks), whereas the distal regions are in an open chromatin structure (red filaments) in the center of the nucleus. Only two chromosomes are depicted (including one acrocentric).

3. Comparative and evolutionary genomics of vertebrates

(i) Several basic structural and functional properties are shared by the genomes of all vertebrates. These properties include a bimodal gene density (see Sections 1 and 2), a biphasic (early-late) replication, a higher recombination rate in telomeric regions of chromosomes, and a higher level of transcription in GC-rich regions.
The genomes of warm- and cold-blooded vertebrates show, however, a major difference, the GC-richest isochores (H2 and H3) of the former having much higher GC levels than those of the latter. In other words, the gene-dense regions of the ancestral genome core of cold-blooded vertebrates are not as GC-rich as those of the genome core of warm-blooded vertebrates (Fig. R10). As a consequence, chromosomes lack clearly defined GC-rich bands. The genomes of cold-blooded vertebrates also differ from those of warm-blooded vertebrates in other properties, such as CpG and methylation levels, CpG islands, etc. (see [iv] below).

(ii) The compositional evolution of vertebrate genomes is characterized by the predominance of a conservative mode; nucleotide substitutions take place, but the compositional patterns do not change (Fig 11.1D). Alternatively, in the transitional (or shifting) mode, nucleotide substitutions are accompanied by compositional changes, the most striking being a GC increase that took place in the ancestral genome core, the gene-dense part of the genome of the cold-blooded vertebrates (Fig. R10). Two other transitional modes have been recognized, the murid shift (Fig. 11.1B) and the whole genome or horizontal shift (see Fig. 11.1C and Part 11).

![Figure R10](11.7). Scheme of the compositional organization of the genomes of cold- and warm-blooded vertebrates. The ancestral genome core of the former is enriched in GC in the genome core of the latter.

(iii) In the conservative mode, a comparison of the average frequencies of codons from orthologous GC-poor and GC-rich genes of species belonging to different mammalian orders, such as calf and human, indicates no difference (Fig. 11.4), and a comparison of orthologous genes shows a high level of compositional similarity in third codon positions (Fig. R11). One should now consider that GC3 similarity is maintained over the whole compositional spectrum, that GC3 is correlated with isochore GC, that extant eutherian orders evolved independently from each other for about 100 million years, and that a strong mutational AT bias exists at least in the human genome (Fig. 11.5), and possibly in all vertebrate genomes. An explanation for the results just described is natural selection, in fact essentially negative (purifying) selection, against average regional compositions going beyond an upper and, especially (because of the mutational AT bias), a lower GC threshold.

In the case of coding sequences and of regulatory sequences, a classical scheme of negative selection on individual changes may be invoked. This can not be applied, however, to noncoding sequences (see Part 12). A possible explanation for the compositional conservation of noncoding sequences is that of critical changes (Fig. R12). As shown in Fig. R12, it is assumed (on the basis of its compositional conservation in different mammalian
or avian orders) that the high average GC level of the GC-rich isochore under consideration (set in Fig. R12 at 55% GC) corresponds to a compositional optimum. It is known that, at least in the human genome, there is a strong mutational AT bias and a deamination of mCpG, hence the higher density of changes (blue arrows) pointing towards AT, compared to that of the changes (red arrows) pointing towards GC. It is assumed that changes leading to average GC levels within the thresholds indicated by the broken horizontal black lines are tolerated and do not change the fitness of the individual. When the accumulation of AT changes is large enough, a few additional "critical changes" (shown as pointing towards AT because of the strong AT bias) may push the average GC beyond the lower

![Diagram of structural and compositional transitions](image)

**Figure R12 (12.3).** A scheme of the conservative mode of compositional evolution of a GC-rich intergenic noncoding region from a warm-blooded vertebrate. Red and blue arrows refer to changes towards GC or AT. Thick arrows refer to critical changes (see text).
threshold and cause a “phase transition” leading to structural changes in DNA and in the chromatin, which affect the expression of genes located within or near the isochore under consideration.

It should be noted that, until regional selection has acted, only neutral and nearly neutral changes will be detected (see Fig. R13). In other words, the neutral and nearly neutral theories account for features of molecular evolution that take place within certain compositional thresholds. They represent, therefore only one mode, the neutral mode, of genome evolution, which is checked by the predominant and pervasive control of selection on compositional patterns, the selective mode of genome evolution.

![Figure R13](12.4). Time course of compositional evolution of a GC-rich noncoding intergenic region from a warm-blooded vertebrate in the conservative mode. It is assumed that the overall compositional trend is dominated by an AT bias. In an early phase, the average GC of the region, initially visualized at its compositional optimum (set at 55% GC), is decreasing but remains within a tolerated range (whose arbitrary thresholds are indicated by the horizontal heavy broken lines). In a late phase, the average GC trespasses the lower threshold, the region undergoes a structural change which is deleterious for gene expression and is subject to negative selection. Until then, the changes are neutral or nearly neutral.

It should be stressed that the drift pushing a region beyond the acceptable thresholds is not only influenced by mutational biases but also by recombination events and transposon insertions. While these factors may slow down, accelerate or alter the direction of the compositional drift, the final control is still exerted by natural selection.

While the stress has been put so far on negative selection, positive selection may also play a role in coding sequences.

In summary, natural selection can explain the maintenance of isochores and is pervasive at the genome level. Natural selection (essentially negative, or purifying, selection) acts both on single nucleotide changes in coding and regulatory sequences, and on regional changes in noncoding sequences (through critical changes).

An obvious implication of selection acting on noncoding sequences is that the latter fulfill a function. Their interpretation as junk DNA can be rejected on several grounds. First of all, on a gross scale, genome size is highly conserved in both mammals and birds. Since noncoding sequences represent 97-98% of mammalian genomes, this means that size changes in intergenic sequences are generally modest, on the average. Second, regulatory sequences are present in noncoding sequences, sometimes dozens of kb away from the
genes that they control. Third, there is increasing evidence for long stretches of sequence conservation in noncoding sequences. Fourth, the expression of integrated viral sequences depends upon the matching or non-matching compositional properties of the flanking host sequences (Fig. R4). Fifth, compositional (isochore) patterns are conserved even when sequence homology is lost, as illustrated by the human/mouse comparison of the MHC locus (Fig. 11.6). In conclusion, genome (and isochore) sizes appear to be under selection. This is also true for the mitochondrial genome, which covers a five-fold size range in eukaryotes.

(iv) If we now consider the transitional, or shifting, mode of evolution (Fig. R10) and compare third codon positions of human (or chicken) genes with orthologous *Xenopus* genes, GC-poor genes are compositionally similar, whereas GC-rich genes show increasing differences (Fig. R14). In contrast, orthologous human/chicken genes are compositionally similar (Fig. 11.10). Likewise, if we compare the average codon frequencies of human (or chicken) genes with orthologous *Xenopus* genes, we find no difference for GC-poor genes and large differences for GC-rich genes (Fig. 11.9), whereas no differences are found in either case for orthologous human and chicken genes (Fig. 11.11). If the emergence of homeothermy was accompanied by an increasing optimal GC level shifting upward the lower threshold of acceptable average composition of the ancestral genome core (corresponding to an upper optimal level), then the compositional transition would strictly require only negative selection, although a contribution by positive selection is likely (Fig. R15). In conclusion, negative selection, far from simply maintaining the status quo (and being “boring”, Kondrashov, 2000), can help genome adaptation (and be “creative”, Schmalhausen, 1949).

The independent compositional transitions that occurred in the ancestral lines of mammals and birds were accompanied by other functionally important changes: 1) CpG and

![Figure R14](11.8). Correlation between C3 and G3 values of orthologous genes from human, or chicken, and *Xenopus*. The orthogonal regression lines are shown together with the diagonals (dashed lines).
methylation levels decreased by a factor of two in mammals/birds compared to fishes/amphibians, but, remarkably, remained stable in each of the two groups (Figs. 11.12 and 11.13), indicating two equilibria separated by a transition; interestingly, reptiles show a wider DNA methylation range (and a wider compositional heterogeneity) compared to the other classes of vertebrates. 2) CpG islands were independently formed at the emergence of mammals and birds; indeed, they are only found in warm-blooded vertebrates (Figs. 5.8-5.10). 3) The nucleotide context of the ATG initiator codons showed an increase in its GC level, at least in mammals (Table 5.1). 4) The formation of a strong banding pattern in warm-blooded vertebrates was paralleled by a higher level of karyotype changes: the disadvantages associated with this increased chromosome instability were apparently compensated by the increased speciation, which, in turn, facilitated the invasion of new ecological niches. 5) Finally, hydrophobicity increased in the encoded proteins (Fig. 12.3), entailing a faster folding of highly expressed proteins.

(v) Natural selection implies the elimination of disadvantages and/or the acquisition of advantages. As far as the compositional transition is concerned, we originally proposed that the selective advantages were the higher thermal stabilities simultaneously achieved at the DNA, RNA and protein levels to cope with the higher body temperature of warm-blooded vertebrates. This can be understood because in a small taxon, such as vertebrates, most factors acting on the compositional pattern of the genome are likely to be similar and because body temperature is the major physiological property that underwent a change at the emergence of warm-blooded vertebrates. The situation of vertebrates is in sharp contrast with that of prokaryotes in which different factors have an input on genome composition of different taxa, making it very difficult, if not impossible, to identify them and assess their influence. However, when closely related bacterial genomes were compared, a higher optimal growth temperature was observed to be accompanied by a higher GC level (Part 12).

The evidence supporting the thermodynamic stability hypothesis comes from the following observations. As far as DNA is concerned, the appearance of GC-rich isochores is correlated with body temperature in vertebrates (see Part 12, Chapter 4). The reasons why only one small part (10-15%) of the genome, the gene-dense part of it, underwent the GC increase in both its coding and noncoding sequences can now be understood. Indeed, in interphase nuclei the genome core is present in very open chromatin structures (Fig. R16), and DNA stabilization may have become necessary when body temperatures increased with the appearance of homeothermy. This stabilization was not required in the genome desert, where the compact chromatin structure could provide the necessary stabilization.

The reasons why only one small part (10-15%) of the genome, the gene-dense part of it,
underwent the GC increase in both its coding and noncoding sequences can now be understood. Indeed, in interphase nuclei the genome core is present in very open chromatin structures (Fig. R16), and DNA stabilization may have become necessary when body temperatures increased with the appearance of homeothermy. This stabilization was not required in the genome desert, where the compact chromatin structure could provide the necessary stabilization.

In the case of RNA, the GC increase leads to the stabilization of stem-and-loop structures against higher temperature.

As far as proteins are concerned, we demonstrated an increase in hydrophobicity with increasing GC₃. This can be understood because quartet and duet codons are positively and negatively correlated with GC₃, respectively, and quartets comprise three abundant hydrophobic amino acids (leucine, alanine and valine), whereas duets comprise six abundant hydrophilic amino acids (lysine, asparagine, glutamine, histidine, glutamic acid and aspartic acid). This increased hydrophobicity was possibly advantageous not only for protein stability but also for increasing the folding rate of the highly expressed proteins that are encoded in the genome core. Finally, as far as the primum movens problem is concerned, clearly the driver was DNA. This is the simplest way to account for the fact that both non-transcribed and non-translated sequences of the ancestral genome core also underwent the change.

At this point, it should be mentioned that several objections have been raised against the
thermodynamic stability hypothesis, but these have been rebutted (see Part 12, Chapter 5). Moreover, a number of alternative hypotheses have been put forward to account for the formation of GC-rich isochores and their maintenance (Table 12.4). While some of them may explain one or two of the reported observations, none can account for most of the other known facts, whereas natural selection can. Indeed, these alternative hypotheses cannot explain why regional compositional changes 1) never occurred in cold-blooded vertebrates living at moderate or low temperatures, 2) were similar in the independent lines of mammals and birds, 3) reached an equilibrium at least 100 million years ago in mammals and were maintained since then, 4) were different in exons, introns, intergenic sequences, and in coding sequences corresponding to different secondary structures of proteins, 5) affected only a small part of the genome (the ancestral genome core) in both coding and noncoding sequences. Other questions could also not be answered, e.g. why the mutational AT bias of the human genome did not lead to the formation of a GC-poor genome, why the increased substitution rates of murids did not increase rather than decrease compositional contrasts, why a decrease in CpG and methylation occurred at the transition from cold- to warm-blooded vertebrates, or why CpG islands were formed and the pre-ATG nucleotides were changed.

4. The eukaryotic genome

Although our investigations were mainly focused on the vertebrate genome, we have explored other eukaryotic genomes, ranging from organelle genomes to protists, invertebrates and plants, and found that some of the basic compositional features present in vertebrate genomes are remarkably widespread.

Indeed, genome compartmentalization was not only found in warm-blooded, and to a smaller extent, in cold-blooded vertebrates, but also in other eukaryotes, ranging from trypanosomes to Drosophila. At least in some cases, this compartmentalization seemed also to be linked to the thermodynamic stability of DNA.

The existence of two gene spaces, the gene-dense (genome core) and the gene-poor (genome desert) regions also appears to be quite widespread. Indeed, such regions were not only found in vertebrates, but also in insects (Drosophila, Anophelus) and plants. Two classes of genes, characterized by different intron sizes and different expression levels, are associated with the two gene spaces, short introns and high expression levels being found in the gene-dense genome core, long introns and low expression levels in the gene-poor genome desert. In compositionally compartmentalized genomes, gene density was correlated with GC levels. At least in vertebrates and plants, gene-dense regions tend to be located near telomeres in chromosomes, and to be characterized by early replication, high recombination and strong transcription.

At first sight, the gene distribution of Gramineae is completely different from that of vertebrates, because genes are concentrated in regions of the genome (collectively called the "gene space") characterized by a very narrow compositional range. This situation is, however, due to a recent massive integration of nested transposons in the gene-dense regions of the genome. These transposons then dictated the overall compositional properties of the invaded regions.
Linear correlations were found between GC (and GC₃) levels of genes and GC levels of the corresponding introns and flanking sequences.

Finally, the mitochondrial genome of yeast taught us several lessons concerning the reversible transconformations undergone by stem-and-loop systems of ori sequences under the effect of temperature, the formation of coding sequences (the var1 gene) from noncoding sequences, the contraction-expansion phenomena of noncoding sequences, the existence of genomes without genes, as well as that of dispensable genomes. As discussed in Part 2, the abundant noncoding sequences of the large mitochondrial DNA from yeast cause genome instability, but this is compensated by a finer modulation of crucial functions, such as replication, transcription and recombination, compared to its compact counterpart in animals.

5. The prokaryotic genome

The concept of genome phenotype can be extended to prokaryotes. In this case, genome composition, although not showing the large internal heterogeneity of many eukaryotic genomes, covers a very wide spectrum. The original explanation for this finding was that genome composition shifts because of directional mutations due to biases in replication enzymes (Freese, 1962; Sueoka, 1962). It was argued, however, that, while these mutational biases are acceptable as the mechanism of the compositional changes, they are not necessarily its cause (Bernardi et al., 1988). The question is whether the resulting overall compositional changes are only determined by the vagaries of random mutations, the mutator mutations, in the genes encoding the protein sub-units of the replication machinery, or are under the control of natural selection. Two facts favour the second conclusion.

The first one is that changes in nucleotide composition correspond to changes in amino acids and, in turn, to changes in the hydrophobicity and in the secondary structure of proteins. If changes only depended upon mutational bias, then one should accept the untenable viewpoint that the corresponding functional changes are selectively irrelevant.

![GC level ranges of DNA from warm- and cold blooded vertebrates, invertebrates, protists and prokaryotes.](R17.png)

Figure R17 (12.19). GC level ranges of DNA from warm- and cold blooded vertebrates, invertebrates, protists and prokaryotes.

The second fact is that GC changes cover a much wider range in the genomes of prokaryotes and unicellular eukaryotes than in those of invertebrates, cold-blooded verte-
brates and warm-blooded vertebrates, these three groups showing a progressively narrower range. Since the potentially relevant mutator mutations in the replication machinery responsible for different mutational biases presumably are comparable in all classes of organisms, and since the spread of GC levels is so different, the only explanation for the different ranges of Fig. R17 is that there is a different selection on the mutations in the replication machinery of these different classes of organisms, some changes being quickly eliminated by negative selection.

What we propose is that genome composition in prokaryotes is under the control of natural selection which optimizes the adaptation of genome composition to the environmental conditions (temperature being only one among many factors). Because of the compactness of the prokaryotic genomes, which are essentially made up of coding and regulatory sequences, natural selection (mainly negative selection) operates on single nucleotide changes, whereas in vertebrate genomes regional changes also play an important role.

Since the different base compositions of prokaryotic genomes were thought to be due to biases in replication enzymes, it was supposed that the same process should also explain the formation and maintenance of isochores in vertebrate genomes. In contrast, the results presented in this book in favour of natural selection as the factor responsible for the isochores of vertebrates and the data of Fig. R17 lead to explaining the different compositions of prokaryotic genomes according to the natural selection scheme proposed here for the isochores, thus reversing a trend that has dominated the literature for the past 40 years.
Conclusions

According to the classical neo-Darwinian selection theory, natural selection, "the preservation of favourable variations and the rejection of injurious variations" (Darwin, 1859), or the differential multiplication of mutant types, essentially occurs through the elimination of the progeny of individual organisms carrying deleterious mutations (negative, or purifying, selection), and only very rarely via the preferential propagation of organisms with advantageous mutations (positive selection).

Twenty years ago, Motoo Kimura formulated the "mutation-random drift theory" according to which "the main cause of evolutionary change at the molecular level – change in the genetic material itself – is random fixation of selectively neutral or nearly neutral mutants". However, "The neutral theory is not antagonistic to the cherished view that evolution of form and function is guided by Darwinian selection, but brings out another facet of the evolutionary process by emphasizing the much greater role of mutation pressure and random drift at the molecular level", in which "increases and decreases in the mutant frequencies are due mainly to chance" (Kimura, 1983).

So far, the classical selection and the neutral theories have been generally visualized as two opposite viewpoints. However, these theories may be seen as complementary. Indeed, one could consider that the selection theory applies, to a large extent, to coding and regulatory sequences, but not to the vast majority of the noncoding sequences that represent 97-98% of a typical mammalian genome. Vice versa, the neutral theory may apply to the vast majority of noncoding sequences, but only partially to coding and regulatory sequences. This picture is, however, only valid on a first approximation. Indeed, both theories only took into consideration single-nucleotide changes. Unfortunately, whether the result of random fixation or selection, single nucleotide changes, cannot account for genome properties such as the genome compartmentalization into isochores, the genome phenotype and the genomic code.

Obviously, a "paradigm shift" was needed to reconcile the indisputable abundance of neutral changes with our results, which inescapably implied natural selection (a conclusion strengthened by sixteen years of critical scrutiny; see Part 12). The first step toward this paradigm shift was made by postulating that changes in base composition could lead to regional phase transitions in DNA/chromatin structure that were deleterious for the expression of the genes located in or near the regions under consideration. This isochores selection proposal (Bernardi et al., 1988) eliminated the need for selection on individual nucleotides in the long non-coding regions of the genome and provided a first model for our previous suggestion (Bernardi and Bernardi, 1986a) that natural selection was the basic explanation for the formation and maintenance of GC-rich isochores. A further step was made here by proposing an improved model in which the accumulation of nucleotide changes (known to be characterized by a strong AT bias) leads to critical changes that, although being single-nucleotide changes, cause the average GC of the region to trespass the tolerated compositional range. The consequence is a transition in DNA/chromatin structure, that is deleterious for the expression of the embedded or flanking genes and affects the fitness of the individual. These compositionally altered regions are then
removed in the progeny by **negative (purifying) selection**, which is also operative, at the individual nucleotide level, in coding and regulatory sequences. This model accounts for the **conservative mode of evolution** of vertebrate genomes. In fact, only negative selection is strictly required also in the **shifting mode of evolution**. In this case, the gene-dense regions of the ancestors of warm-blooded vertebrates increased their GC level in response to the upward shift of the compositional optimum associated with the emergence of homeothermy and the consequent need for stabilizing thermodynamically the DNA present in the open chromatin of the gene-dense isochores.

Our DNA/chromatin transition model, which we propose to call the **neo-selectionist model**, has the merit of reconciling the prevalence in numbers of neutral, nearly neutral and slightly deleterious mutations with the pervasive control exerted at the regional level by natural selection (essentially negative selection). If one looks at a broader perspective, this “reconciliation” should not hide, however, the fact that the neutral theory and the neo-selectionist model proposed here lead to **two very different general views of genome evolution**. This contrast can be best visualized by focusing on three points.

(i) Kimura realized that his neutral theory raised a dilemma, “*why natural selection is so prevalent at the phenotype level and yet random fixation of selectively neutral and nearly neutral alleles prevails at the molecular level*”. Although Kimura showed a remarkable insight by stating “*The answer to this question, I think, comes from the fact that the most common type of natural selection at the phenotype level is stabilizing selection*”, he concluded that “*laws governing molecular evolution are clearly different from those governing phenotypic evolution*”. A beginning to the solution of this serious dichotomy was resolved by our proposal of 1986 that natural selection controls not only the “**classical phenotype**” of form and function (or, in molecular terms, of the proteins and of their expression), as generally accepted, but also the **genome phenotype**, the compositional properties of the genome and all their functional implications. In other words, our proposal is that **natural selection extends to the molecular level, to the genome itself, and to its compositional patterns** (and not only to the single nucleotides of coding or regulatory sequences). This natural selection operating on the genome phenotype could be called **compositional selection** (which is regional in vertebrates, but typically genome-wide in prokaryotes), and would act in addition to the **classical selection**, namely natural selection operating on the classical phenotype, and to the neutral changes.

(ii) When the compositional selection has not yet acted, obviously only neutral and nearly neutral changes will be detected. This is to say that the neutral and nearly neutral theories account for features of molecular evolution which take place within certain compositional thresholds, and represent one mode, the **neutral mode**, of genome evolution. This neutral mode is, however, dominated by the pervasive control of natural selection in the **selection mode**. As already mentioned, the overall model could be called the **neo-selectionist model** of evolution.

(iii) Kimura’s view that “*increases and decreases in the mutant frequencies are due mainly to chance*” led him, logically, to the idea of the “**survival of the luckiest**”. Obviously, the pervasive control of natural selection on the “genome phenotype” not only solves Kimura’s dilemma, but leads us back to the “**survival of the fittest**”. Although these expressions have been criticized as tautologous, they summarize very different assessments of the role of chance in evolution. It is then difficult not to conclude that, **even at the molecular**
level, “it may be said that natural selection is daily and hourly scrutinizing, throughout the world, every variation, even the slightest; rejecting that which is bad, preserving and adding up all that is good; silently and insensibly working, whenever and wherever opportunity offers, at the improvement of each organic being in relation to its organic and inorganic conditions of life” (Darwin, 1859).
Abbreviations and acronyms

- A, CsCl band asymmetry, $<\rho>-\rho_0$
- BAC(s), Bacterial Artificial Chromosome(s)
- BAMID, bis(acetomercurimethyl)dioxane
- BLV, Bovine Leukaemia Virus
- bp, base pairs(s)
- cM, centimorgan(s)
- C₀t, initial DNA concentration x reannealing time
- cox 1, citochrome oxidase sub-unit 1
- c value, genome size
- cyt b, apocytochrome b
- DAPI, 4,6 diamino-2-phenyindole
- DNase, deoxyribonuclease
- ESAG(s), Expression Site Associated Gene(s)
- EST(s), Expressed Sequence Tag(s)
- FITC, fluorescein isothiocyanate
- Gb, gigabase(s), or billions of base pairs
- GC, molar ratio of guanosine and cytidine in DNA
- GC₁, GC₂, GC₃, GC levels in first, second and third codon position
- GTG bands, G bands by trypsin using Giemsa
- H, intermolecular compositional heterogeneity
- HBV, Hepatitis B Virus
- HIV-1, Human Immunodeficiency Virus 1
- HTLV-1, Human T-cell Leukaemia Virus 1
- IRES, Internal Ribosome Entry Site
- kb, kilobase(s), or thousand(s) of base pairs
- LINE(s), Long Interspersed Repeat(s)
- LTR(s), Long Terminal Repeat(s)
- Mb, Megabase(s), or millions of base pairs
- MHC, Major Histocompatibility locus
- MMTV, Mouse Mammary Tumour Virus
- MoMLV, Moloney Murine Leukemia Virus
- MuLV, Murine Leukemia Virus
- Mya, Million(s) years ago
- Myr(s), Million year(s)
- N(s), undetermined nucleotide(s)
- ORF, Open Reading Frames
- ori, origin of DNA replication
- PCR, Polynucleotide Chain Reaction
- pg, picogram(s)
- PKR, Particulate Protein Kinase
- rf, ligand/nucleotide molar ratio
• **RIDGE(s)**, Region(s) of Increased Gene Expression
• **RSV**, Rous Sarcoma Virus
• **RFLP(s)**, Restriction Fragment Length Polymorphism(s)
• **SCC**, Sequence Compositional Complexity
• **sd**, standard deviation
• **SINE(s)**, Short Interspersed Repeat(s)
• **SNP(s)**, Single Nucleotide Polymorphism(s)
• **SV40**, Simian virus 40
• **Tb**, body temperature
• **Tmax**, maximal Tb sustained for extended periods of time
• **Topt**, optimal growth temperature
• **T-DNA**, Transferred DNA (from *Agrobacterium* plasmids)
• **TE**, Transposable Elements
• **TRITC**, tetra methyl rhodamine (3 isothiocyanate)
• **UTR(s)**, Untranslated Sequence(s)
• **YAC(s)**, Yeast Artificial Chromosome(s)
• **5mC**, 5-methylcytosine
• **<p>**, mean buoyant density
• **ρo**, modal buoyant density
References

References are given in alphabetical order for papers with one or two authors, in chronological order for the papers with more than two authors. Only first authors are quoted for sequence papers.


Brosius J. (1999) Genomes were forged by massive bombardments with retroelements and retrosequences. Genetica 107: 209-238.


de Miranda A.B., Alvarez-Valin F., Jabbari K., Degrate W.M., Bernardi G. (2000). Gene expression, amino acid conservation, and hydrophobicity are the main factors shaping
codon preferences in *Mycobacterium tuberculosis* and *Mycobacterium leprae*. J. Mol. Evol. 50: 45-55


zation of HTLV-1 and HIV-1 proviral sequences in chromosomes of persistently infected cells. Chromosome Res. 7: 177-183.


Ikemura T. (1982). Correlation between the abundance of yeast transfer RNAs and the
occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. J. Mol. Biol. 158: 573-597.


provirus in H-19 hasmter tumor cell line is integrated adjacent to the negative regulatory region. Gene 174: 9-17.


organization as revealed by the use of cytosine methylation-sensitive restriction endonucleases. Genomics 15: 472-482.


Pradet-Balade B., Boulmé F., Beug H., Müllner E.W., Garcia-Sanz J.A. (2001). Transla-


classes of middle-repetitive DNA in the *Arabidopsis thaliana* genome. Nucleic Acids Res. 24: 3017-3022.


An update (May 2005)

Part 2. Lessons from a small dispensable genome, the mitochondrial genome of yeast
A review paper (Bernardi, 2005) discusses the origins of replication of the mitochondrial genome of yeast in the light of some recent proposals.

Part 3. The organization of the vertebrate genome
Misunderstandings about isochores are not finished, as witnessed by a paper by Cohen et al. (2005). An explanation of the reasons that led these authors to wrong conclusions was given by Clay and Bernardi (2005).

Part 5. Sequence distribution in the vertebrate genomes
Investigations on the methylation levels of vertebrate genomes (Varriale et al., 2005) have confirmed the higher levels (about 2-fold) previously reported by Jabbari et al. (1997) of fishes/amphibians compared to mammals/birds. They have shown, in addition, that genomes from polar fish species have a higher methylation level compared to fish from tropical/temperate seas/oceans. This stresses the inverse correlation between DNA methylation and body temperature. The results of Varriale et al. (2005) have also shown that reptilian genomes cover the whole spectrum of methylation levels comprised between fishes/amphibians and birds/mammals. This can be understood in terms of the different average body temperatures of different reptiles.

Part 6. The distribution of integrated viral sequences, transposons and duplicated genes in the mammalian genome
A further analysis of duplicated genes in the human genome (in terms of introns, repeats and associated CpG islands) is in press (Rayko et al., 2005).

Part 7. The organization of chromosomes in vertebrates
A detailed analysis of human chromosomal bands (Costantini et al., 2005) has generalized the previous conclusion (Saccone et al., 2001) that Giemsa and Reverse bands are due to their compositional contrast with flanking bands: these are GC-richer in the case of Giemsa bands and GC-poorer in the case of Reverse bands. Our conclusions concerning the open chromatin structure of GC-rich, gene-rich bands and central location in the interphase nucleus (as opposed to the closed chromatin structure and peripheral location of GC-poor, gene-poor bands; Saccone et al., 2002) has been confirmed by another laboratory (Gilbert et al., 2004).

Part 12. Natural selection and genetic drift in genome evolution: the neo-selectionist model
Our results on the dependence of genomic GC levels upon optimal growth rate of prokaryotes (Musto et al., 2004) have been criticized as not robust (Marashi and Ghalanber, 2004). We have shown, however (Musto et al., 2005), that the latter conclusion was based on a wrong interpretation of our data.
References


