cient quantities of extremely pure polypeptides to be used directly for the generation of both monoclonal and polyclonal antibodies. Under certain circumstances, individual CBB-stained protein spots have been excised from whole gels and used for the direct immunization of rabbits for polyclonal antibody formation.

In addition to antibody production, recent analytical advances in the area of Edman N-terminal amino acid microsequencing have made it possible to obtain N-terminal and internal amino acid microsequence information for polypeptide spots isolated directly from 2D-PAGE gels, as schematically illustrated in Figure 5.

Following 2D-PAGE, proteins are electrophoretically transferred to PVDF membrane supports and visualized using either CBB or Ponceau S staining. CBB staining is two to three times more sensitive that Ponceau S staining, but Ponceau S is more desirable if N-terminal blockage necessitates subsequent enzymatic hydrolysis of individual proteins. Regardless of the mode of visualization, individual protein spots of interest are then cut from the stained membranes and inserted directly into a gas phase microsequencer and N-terminal amino acid sequencing is performed.

Figure 7 is a 2D-PAGE map of RLE "nuclear sap" proteins electroblotted to PVDF membrane and stained with Ponceau S. If a protein is not N-terminally blocked (e.g., acetylated, glycosylated, etc.), proteins expressed as little as 5–10 pmol (250–500 ng of 50 kDa protein) can be easily sequenced. N-Terminally blocked proteins require additional chemical and/or enzymatic cleavage and purification by high performance liquid chromatography prior to sequence analysis. Under optimal conditions 10–15 amino acid residues can usually be obtained from as little as 1–10 pmol of unblocked protein (Figure 7: proteins 392 and 1332), with some proteins such as heat shock protein 60 yielding up to 25 residues. The sequence data can then be used to search various protein sequence databases (e.g., PIR, PATCHX, SWISSPOT) for protein identification or sequence homology comparisons. If a protein has not been characterized or its identity is unknown (Figure 7: proteins 392, 1079, 1332), the partial sequences can be used for additional characterization studies. These include the large-scale chemical synthesis of corresponding poly(oligo)-peptides for both polyclonal and monoclonal antibody preparation as well as for the design and synthesis of specific oligonucleotide probes for cDNA isolation, gene cloning, and subsequent genetic analysis.

5 CONCLUDING REMARKS

The technique of 2D-PAGE provides an extremely powerful analytical method for the separation of complex protein mixtures. With recent refinements in protein microsequencing techniques, it provides a valuable link between protein biochemistry and molecular biology.

See also HPLC OF BIOLOGICAL MACROMOLECULES; PROTEIN ARCHITECTURE AND ANALYSIS; PROTEIN PURIFICATION; PROTEINS AND PEPTIDES, ISOLATION FOR SEQUENCE ANALYSIS OF.

Bibliography

Celis, J. E., and Bravo, R. (1984) *Two-Dimensional Electrophoresis of Proteins: Methods and Applications.* Academic Press, New York.

Dunbar, B. S. (1987) *Two-Dimensional Electrophoresis and Immunological Techniques.* Plenum Press, New York.

Hames, B. D., and Rickwood, D. (1990) *Gel Electrophoresis of Proteins: A Practical Approach.* IRL Press, New York.

Luo, L-di and Wirth, P. J. (1993) Consecutive silver staining and autoradiography of ^{35}S and ^{32}P-labeled cellular proteins: Applications for the analysis of signal transducing pathways. *Electrophoresis* 14, 127–136.

Pharmacia LKB Biotechnology AB. (1993) *2-D Electrophoresis Protocol Using IPG Immoboline DryStrips: Instruction Manual.* Pharmacia, Uppsala, Sweden.

Wirth, P. J., Luo, L-di, Benjamin, T., Hoang, T. N., Olson, A. D. and Parmalee, D. C. (1993) The rat liver epithelial (RLE) cell nuclear protein database. *Electrophoresis* 14, 1199–1215.

GENE DISTRIBUTION IN THE HUMAN GENOME

Giorgio Bernardi

Key Words

Chromosomal Bands During mitosis, chromosomes condense and, at metaphase, they are characterized by specific staining properties. Under standard conditions, Giemsa staining produces a total of about 400 bands that comprise, on average, 7.5 million base pairs (Mb) of DNA.

Genome Every living organism contains, in its genome, all the genetic information that is required to produce its proteins (as well as some ribonucleic acids, which are not translated into protein) and is transmitted to its progeny. The genome consists of DNA, which is made up of two complementary strands wound around each other to form a double helix. The building blocks of each DNA strand are deoxyribonucleotides. These are formed by a phosphate ester of deoxyribose (a sugar), linked to one of four bases: two purines [adenine (A) and guanine (G)] and two pyrimidines [thymine (T) and cytosine (C)]. In the DNA double helix, purines pair with pyrimidines (A with T, G with C).

Isochores Long regions of DNA characterized by homogeneity of base composition. Isochore size is 0.2–1.5 Mb or more. Isochores belong to a small number of families having distinct compositions.

This article deals with the organization of nucleotide sequences in the human genome and with the evolutionary history of such organization. Far from being an ensemble of genes scattered over vast expanses or intergenic sequences, the genome is highly ordered from the nucleotide level to the chromosomal level. Nucleotide sequences, whether in the 3% of the genome formed by coding sequences or in the 97% formed by noncoding sequences, obey precise rules that amount to a genomic code.

1 INTRODUCTION

The term *genome* was coined three-quarters of a century ago by the German botanist Winkler, to designate the haploid chromosome set. While current textbooks of molecular biology do not yet go beyond the purely operational definition (i.e., the eukaryotic genome is the sum total of genes and intergenic sequences), a number of molecular biologists have been thinking for some time that the genome is more than the sum of its parts. This implies the existence of structural and functional interactions between the coding sequences (the minority) and the noncoding sequences (the great majority). This general rather vague concept has been changed into a precise one by the discovery, in our laboratory, of compositional genome properties. These properties, which have mainly been defined by investigations on the nuclear genome of vertebrates, are briefly outlined here. They comprise the isochore organization, the compositional patterns of DNA fragments (or molecules) and of coding sequences, the compositional correlations between coding and noncoding sequences and, above all, the gene distribution and its associated functional properties.

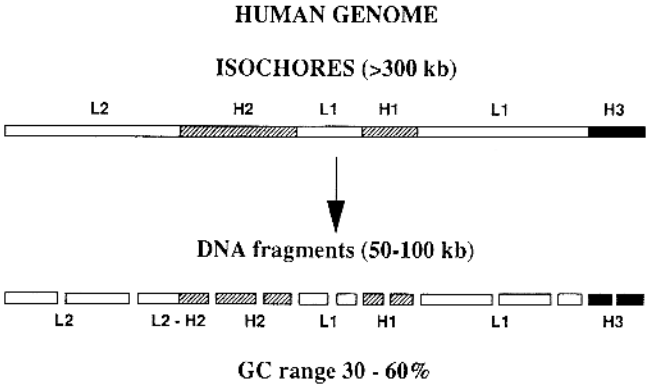

Figure 1. Scheme of the isochore organization of the human genome. This genome, which is a typical mammalian genome, is a mosaic of large (> 300 kb) DNA segments, the isochores. These are compositionally homogeneous (above a size of 3 kb) and can be subdivided into a small number of families: GC-poor (L1 and L2), GC-rich (H1) and (H2), and very GC-rich (H3). The GC range of the isochores from the human genome is 30–60%. [From Bernardi (1993).]

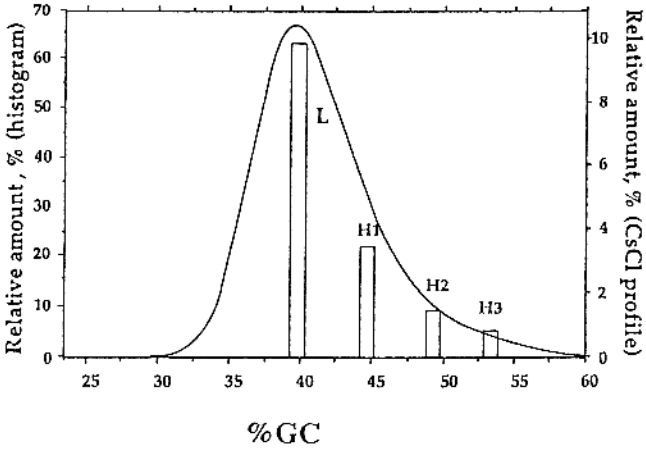

Figure 2. Histogram of the isochore families from the human genome. The relative amounts of major DNA components derived from isochore families L (i.e., L1 + L2), H1, H2, and H3 are superimposed on the cesium chloride profile of human DNA. [From Mouchiroud et al. (1991).]

2 GENOME ORGANIZATION AND GENE DISTRIBUTION

The mammalian genomes are mosaics of *isochores* (Figure 1)—long (> 300 kb) DNA segments that are homogeneous in base composition and cover a 30–60% GC range. This is an extremely wide

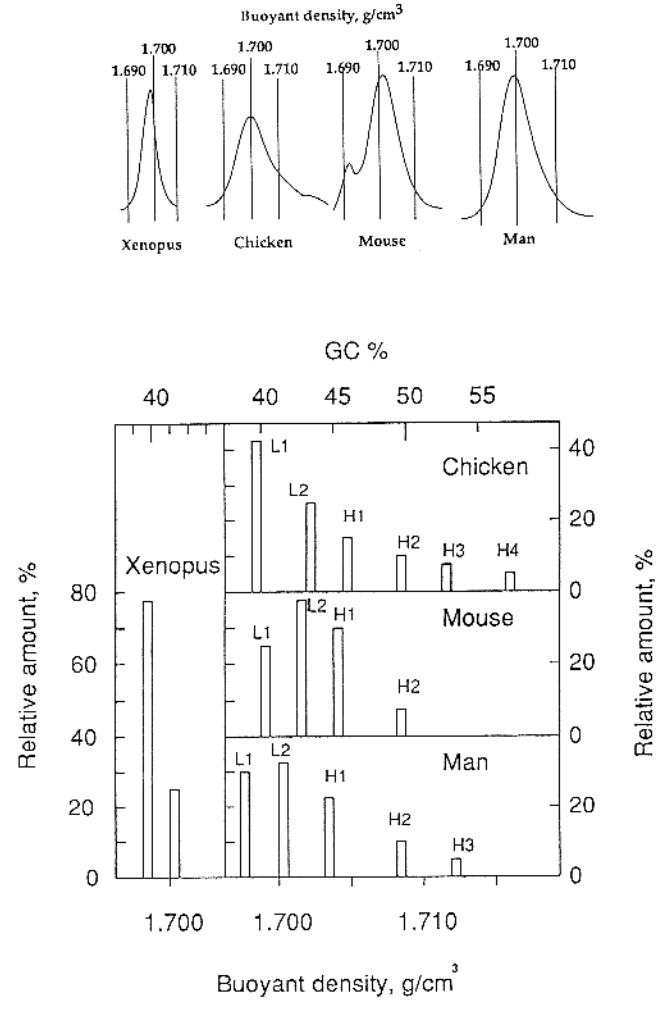

Figure 3. Compositional patterns of vertebrate genomes. *Top:* Cesium chloride (CsCl) profiles of DNAs from *Xenopus*, chicken, mouse, and man. [From Thiery et al. (1976)]. *Bottom:* Histograms showing the relative amounts, modal buoyant densities, and GC levels of the major DNA components from *Xenopus*, chicken, mouse, and man, as estimated after fractionation of DNA by preparative density gradient centrifugation in the presence of a sequence-specific DNA ligand [Ag^+ or bis(acetatomercurimethyl) dioxane) (BAMD)]. The major DNA components are the families of large DNA fragments (see Figure 1) derived from different isochore families. Satellite and minor DNA components (such as ribosomal DNA) are not shown in these histograms. (From Bernardi (1993).]

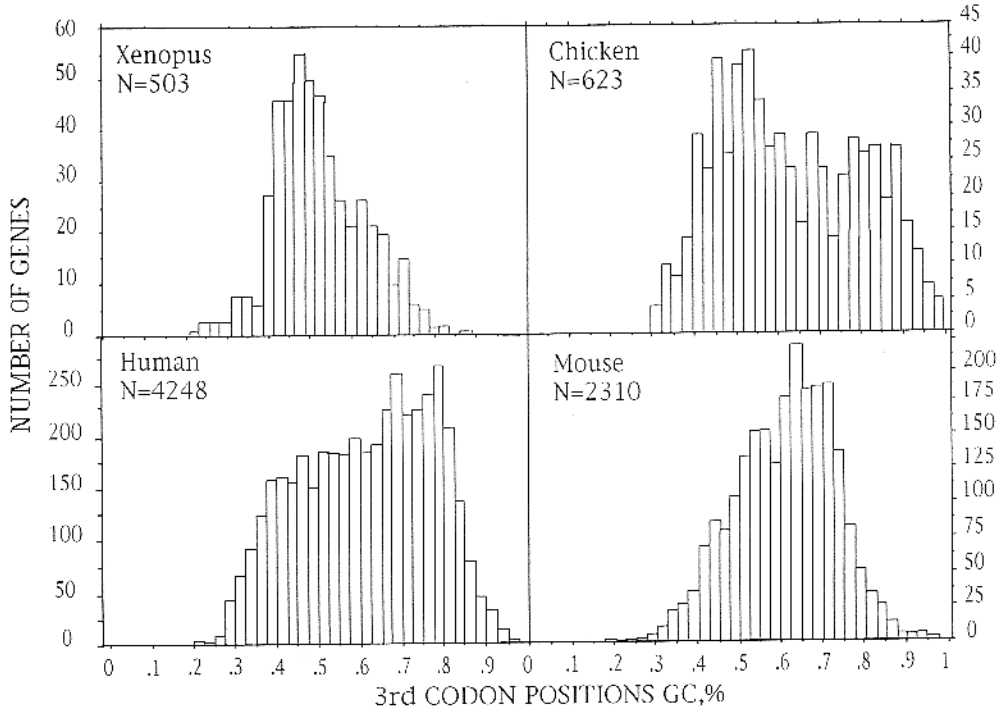

Figure 4. Compositional distribution of third-codon positions from vertebrate genes. The number of genes taken into account is indicated. A 2.5% GC window was used. [From Bernardi (1993).]

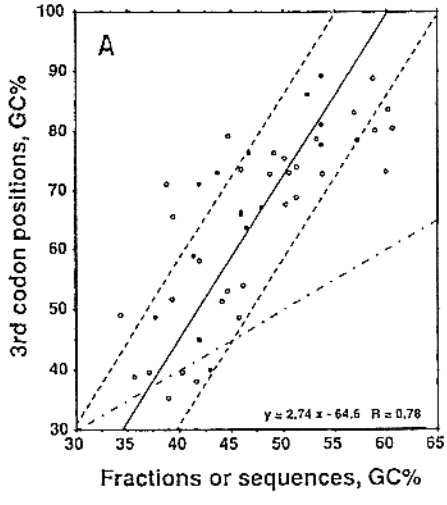

Figure 5. (A) GC levels of third-codon positions from human genes are plotted against the GC levels of DNA fractions (dots) or extended sequences (circles) in which the genes are located. The correlation coefficient and slope are indicated. The dot–dash line is the diagonal line (slope = 1). GC levels of third-codon positions would fall on this line if they were identical to GC levels of surrounding DNA. The broken lines indicate a ±5% GC range around the slope. [From Mouchiroud et al. (1991).]

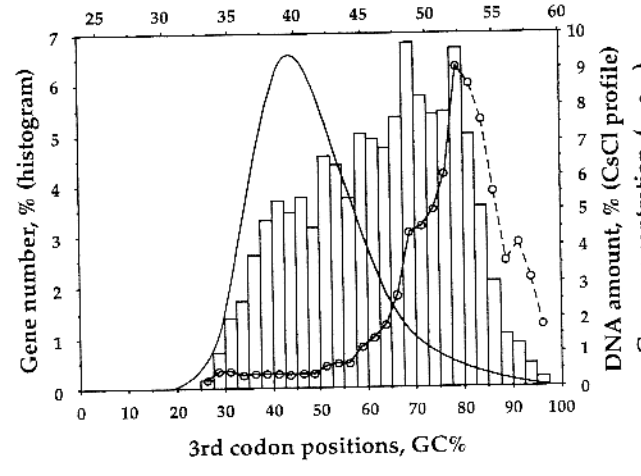

Figure 6. Profile of gene concentration in the human genome as obtained by dividing the relative amounts of genes in each 2.5% GC interval of the histogram by the corresponding relative amounts of DNA deduced from the CsCl profile. The apparent decrease in gene concentration for very high GC values (broken line) is due to the presence of rDNA in that region. The last concentration values are uncertain because they correspond to very low amounts of DNA. [From G. Bernardi (1993).]

range, almost as wide as that covered by all bacterial genomes (25–72% GC). In the human genome, isochores can be assigned to two GC-poor families (L1 and L2), representing two-thirds of the genome, and to three GC-rich families (H1, H2, and H3), forming the remaining third (Figure 2).

The *compositional distributions* of large (> 100 kb) genome fragments, such as those forming routine DNA preparations, of exons (and particularly of their third-codon positions) and of introns, represent *compositional patterns*. These correspond to *genome phenotypes* in that they differ characteristically not only between cold- and warm-blooded vertebrates, but also between mammals and birds and even between murids and most other mammals (Figures 3 and 4).

Compositional correlations exist between exons (and their codon positions) and isochores (Figure 5), as well as between exons and introns. These correlations concern, therefore, coding and noncoding sequences and are not trivial, since coding sequences make up only about 3% of the genome, whereas noncoding sequences correspond to 97% of the genome. The compositional correlations represent a *genomic code*. It should be noted that a *universal correlation* holds among GC levels of codon positions (third positions against first and/or second positions). Both the genomic code and the universal correlation are apparently due to compositional constraints working in the same direction (toward GC or AT), although to different extents on coding and noncoding sequences as well as on different codon positions.

The compositional correlations between GC_3 (the GC level of third-codon positions) and isochore GC have a practical interest in that they allowed us to position the coding sequence histograms of Figure 4 relative to the cesium chloride profile of Figure 3 and to assess the *gene distribution* in the human genome. In fact, if one divides the relative number of genes per histogram bar by the corresponding relative amount of DNA, one can see that gene concentration is low and constant in GC-poor isochores L1 and L2, increases with increasing GC in isochore families H1 and H2, and reaches a maximum in isochore family H3, which exhibits at least a 20-fold higher gene concentration compared to GC-poor isochores (Figure 6).

The H3 isochore family has been called the *human genome core* because it corresponds to the functionally most significant part of the human genome. Indeed, the H3 isochore family is endowed not only with the highest gene (and CpG island) concentration, but also with an open chromatin structure (as witnessed by the accessibility to DNases, as well as by the scarcity of histone H1, the acetylation of histones H3 and H4, and the wider nucleosome spacing), with the highest transcription and recombination levels and with an early replication timing. The genes of the genome core have the highest GC_3 levels relative to their flanking sequences, have the shortest exons and introns, exhibit an extreme codon usage, and encode proteins characterized by amino acid frequencies differing from those of proteins encoded by GC-poor isochores.

The human genome core is located in T (telomeric) bands which are essentially formed by GC-rich isochores (mainly of the H2 and H3 families). In contrast, R′ bands—that is, the subset of R bands comprising reverse bands exclusive of T bands, consist of both GC-rich isochores (of the H1 family) and GC-poor isochores. Finally, G(iemsa) bands are formed almost exclusively by GC-poor isochores. The difference in GC level between G bands and T bands is about 15%. About 20% of genes are present in G bands and about 80% in R bands (60% of them in T bands). The location of a majority of genes in T bands is of interest in view of the association of telomeres with the nuclear matrix and envelope.

3 GENE DISTRIBUTION IN VERTEBRATE EVOLUTION

It should be stressed that the gene distribution reported for the human genome seems to have been conserved in evolution, since genes show their highest concentration in the GC-richest isochores of all vertebrates.

In the case of homologous mammalian genes, it has been possible to show that third-codon-position synonymous substitutions exhibit frequencies and compositions that strongly suggest natural selection. Under these circumstances, the compositional changes in noncoding sequences, which are correlated with those occurring in third-codon positions, suggest that noncoding sequences are not junk DNA, but must fulfill some functional role.

As already mentioned, the compositional pattern of the human genome, which is typical of the genomes of most mammals and similar to the genomes of birds, is strikingly different from the compositional patterns of cold-blooded vertebrates, which exhibit a much lower degree of heterogeneity and are characterized by metaphase chromosomes, which do not show R-banding. These different genome phenotypes of warm- versus cold-blooded vertebrates are due to compositional changes. While the gene-poor, GC-poor isochores have undergone little or no compositional change in vertebrates genomes, the gene-rich, GC-rich isochores are those that underwent compositional changes in evolution.

See also MAMMALIAN GENOME; MOUSE GENOME; TANDEMLY REPEATED NON-CODING DNA SEQUENCES.

Bibliography

Aïssani, B., D'Onofrio, G., Mouchiroud, D., Gardiner, K., Gautier, C. and Bernardi, G. (1991) The compositional properties of human genes. *J. Mol. Evol.* 32:497–503.

Bernardi, G. (1995) The human genome: Organization and evolution. *Ann. Rev. Gen.* 29:445–476.

———, and Bernardi, G. (1986) Compositional constraints and genome evolution. *J. Mol. Evol.* 24:1–11.

———, Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., and Rodier, F. (1985) The mosaic genome of warm-blooded vertebrates. *Science,* 228:953–958.

Cacciò, S., Perani, P., Saccone, S., Kadi, F., and Bernardi, G. (1995) Single-copy sequence homology among the GC-richest isochores of the genomes from warm-blooded vertebrates. *J. Mol. Evol.* 39:331–339.

de Lange, T. (1992) Human telomeres are attached to the nuclear matrix. *EMBO J.* 11:717–724.

D'Onofrio, G., Mouchiroud, D., Aïssani, B., Gautier, C., and Bernardi, G. (1991) Correlations between the compositional properties of human genes, codon usage and amino acid composition of proteins. *J. Mol. Evol.* 32:504–510.

Duret, L., Mouchiroud, D., and Gautier, C. (1995) Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J. Mol. Evol.* 40:308–317.

Macaya, G., Thiery, J. P., and Bernardi, G. (1976) An approach to the organization of eukaryotic genomes at a macromolecular level. *J. Mol. Biol.* 108:237–254.

Mouchiroud, D., Fichant, G., and Bernardi, G. (1987) Compositional compartmentalization and gene composition in the genome of vertebrates. *J. Mol. Evol.* 26:198–204.

———, D'Onofrio, G., Aïssani, B., Macaya, G., Gautier, C., and Bernardi, G. (1991) The distribution of genes in the human genome. *Gene,* 100:181–187.

Saccone, S., De Sario, A., Della Valle, G., and Bernardi, G. (1992) The highest gene concentrations in the human genome are in T-bands of metaphase chromosomes. *Proc. Natl. Acad. Sci. U.S.A.* 89:4913–4917.

———, ———, Wiegant, J., Rap, A. K., Della Valle, G., and Bernardi, G. (1993) Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* 90:11929–11933.

Tazi, J., and Bird, A. (1991) Alternative chromatin structure at CpG islands. *Cell,* 60:909–920.

Thiery, J. P., Macaya, G., and Bernardi, G. (1976) An analysis of eukaryotic genomes by density gradient centrifugation. *J. Mol. Biol.* 108: 219–235.

Winkler, H. (1920) *Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreich.* Fischer, Jena, 1920.

Zoubak, S., D'Onofrio, G., Cacciò, S., Bernardi, G., and Bernardi, G. (1995) Specific compositional patterns of synonymous positions in homologous mammalian genes. *J. Mol. Evol.* 40:293–307.

Gene Expression, Regulation of

Göran Akusjärvi

Key Words

Epigenetic Modification Changes in the phenotype that are not due to alterations in the genotype (i.e., mutations in the DNA).

Exons Eukaryotic genes are encoded in discontinuous segments, where the coding portions are interrupted by noncoding sequences of unknown function (see **introns**). Both exonic and intronic sequences are transcribed into a nuclear precursor RNA. The segments of a eukaryotic gene that is preserved in the mature mRNA are the exon sequences. Prokaryotic genes usually are not split and, thus, are encoded by a contiguous DNA sequence.

Introns Represent a segment of DNA that is transcribed into RNA in the nucleus of the eukaryotic cell, but is excised by RNA splicing before the exportation of mature mRNA to the cytoplasm.

Nucleosome The basic structural subunit used to condense DNA in a cell. The nucleosome consists of approximately 200 base pairs of DNA wrapped around a protein core made up of histone proteins.

Promoter The nucleotide sequence in the DNA to which the RNA polymerase binds when it begins transcription.

RNA Splicing The process by which introns are removed from a nuclear precursor RNA during formation of a functional mRNA.

TATA Box A conserved TATAAA sequence found about 25–30 base pairs upstream of the transcription initiation site in eukaryotic RNA polymerase II promoters. (A similar sequence element is also found in prokaryotic promoters, the −10 element.) The TATA box binds the general transcription factor TFIID and helps position the RNA polymerase for correct initiation.

TFIID Transcription factor D for RNA polymerase II: a general transcription factor that interacts with the TATA box. It consists of the TATA-binding protein (TBP), which makes contact with the TATA box and several TBP-associated factors (TAFs), which are required for regulation of transcription.

Translation The process by which the ribosome reads the nucleotide sequence of the mRNA and directs the incorporation amino acids into protein.

U snRNP The uridine-rich, small nuclear ribonucleoprotein particles contained in large quantity in the nuclei of eukaryotic cells. The U1, U2, U4, U5, and U6 snRNPs have all been shown to be involved in RNA splicing. Other U snRNPs serve other functions in the cell.

Genetic information is transmitted between generations of a species in the form of a stable DNA molecule. This molecule is replicated before cell division to ensure that all offspring receive the same genetic constitution. Expression of the genetic information has been summarized in the so-called central dogma, according to which the flow of genetic information in a cell is transmitted from the DNA to an RNA intermediate to a protein.

Organisms are divided into two major groups, depending on whether their cells possess a nucleus: the prokaryotes, which in-