BIOINFORMATICS FOR BEGINNERS

Genes, Genomes, Molecular Evolution, Databases and Analytical Tools

Supratim Choudhuri

With contribution from Dr Michael Kotewicz on the Optical Mapping of DNA

Center for Food Safety and Applied Nutrition, FDA, College Park, Maryland

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
32 Jamestown Road, London NW1 7BY, UK
225 Wyman Street, Waltham, MA 02451, USA
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA

2014 Published by Elsevier Inc.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-12-410471-6

14 15 16 17 18 10 9 8 7 6 5 4 3 2 1

To my Family

Preface

As the title of the book suggests, this book is indeed for "beginners." It is not intended for advanced students of bioinformatics or practicing bioinformaticians. This book has been written from the perspective of an end-user who wants to use the freely available web-based databases and tools for bioinformatic analysis. The audience of this book could include any scientist or student who has a background in basic molecular biology but has not used web-based databases and tools for sequence analysis, or has not done bioinformatic analysis on a regular basis. The total number of chapters is only nine. This is because related sections have been combined into one chapter for coherence and understanding. These sections could have been easily split into separate stand-alone chapters to increase the number of chapters.

More than a decade into the first human genome sequencing, the use of bioinformatic analysis has been steadily increasing. There are more web-based freely available databases and analytical tools than ever before. Modern biology has pervaded even the social sciences. For example, sociologists and psychologists are now probing how the epigenomic effects of environmental factors (including social factors) might shape the personality and behavior of the offspring postnatally. The National Center for Biotechnology Information has established an epigenomics database, which will be immensely useful to scientists in the near future. Thus, bioinformatics has been slowly but steadily pervading all branches of biology and beyond. In keeping with this, more and more bioinformatics books are being written for experts, which do not necessarily cater to the needs of the non-experts.

Because this book is about bioinformatic analysis using web-based databases and tools, the emphasis is on sequence analysis. Global gene-expression profiling has not been emphasized other than a short discussion. The makers of gene-expression analysis platforms provide necessary software for analysis. Lastly, it is not possible to show every type of analysis in a book with a defined word count; nor is it possible to discuss all the links and all the functions associated with a database or analysis. Therefore, this book should serve as an initial guide, and it is expected that the reader will take it upon himself/herself to explore further using the databases and tools. Terms such as program, tool, algorithm, and web server have been used interchangeably throughout the book. These terms essentially mean the same thing in the context of this book. However, the term web server could be used to mean both the hardware and the software.

Because the principal audience of the book is supposed to be non-specialists, it was felt necessary to introduce the science and some core concepts of genomics as well as some important genomic techniques before embarking on the bioinformatic analysis. By the same token, some fundamental aspects of molecular evolution have been discussed in this book because the goal of many applications of bioinformatics is to trace the signatures of molecular evolution, as well as study the relatedness of taxa. In order to minimize the number of references in the text, reviews are cited wherever possible.

Supratim Choudhuri

Acknowledgment

The author would like to acknowledge the invaluable contributions of all scientists and engineers who developed databases and online tools for analysis, and made them freely available. The author would also like to acknowledge the contributions of the groups/institutions/organizations for hosting and maintaining these resources on web servers. A number of links for freely available databases and web-based tools for analysis have been provided throughout the book. Wherever possible, the latest relevant publications (which usually include the previous publications as well) describing these resources have been cited to acknowledge the contribution. The scientific community is truly grateful to the developers of these tools and databases and for making them freely available to facilitate bioinformatic analysis and learning.

The author would like to thank Dr Steve Gendel for his careful reading of the allergenicity prediction section in Chapter 8, and providing helpful suggestions.

The author would also like to thank many colleagues for their encouragement, enthusiasm, and support for the project.

Last but not the least, the author is grateful to Mr Graham Nisbet and Ms Catherine Mullane of Elsevier for making this project a reality, helping to bring it to successful completion, and being available whenever help and advice were needed.

CHAPTER

1

Fundamentals of Genes and Genomes*

OUTLINE

*The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government.

1.1 BIOLOGICAL MACROMOLECULES, GENOMICS, AND BIOINFORMATICS

Genetic information is stored in the cell in the form of biological macromolecules, such as nucleic acids and proteins. The genetic information not only drives the functioning of the whole organism, but also drives the evolutionary engine. Thus, an understanding of the molecular basis of life is fundamental to understanding how genetic information shapes life and drives its evolution. The following discussion captures some fundamental aspects of the structure and function of genes and genomes with special notes (in boxes) on the applications of this information.

1.2 DNA AS THE UNIVERSAL GENETIC MATERIAL

With some exceptions, deoxyribonucleic acid (DNA) is the universal genetic material. In some viruses, termed RNA viruses, RNA is the genetic material. The term **ribovirus** is used for viruses with single- and double-stranded RNA genomes, including retroviruses, which are RNA-based for a portion of their life cycle.[1]

Among the RNA viruses, **retroviruses** are well known; they include the notorious AIDS virus. Retroviruses are unique because in their life cycle they have both RNA and DNA versions of their genome. A complete retrovirus contains an RNA genome. The RNA genome encodes some protein products that are necessary for converting the single-stranded RNA genome into a double-stranded DNA genome and then its subsequent integration into the host genome. One such protein product of the retroviral genome is the reverse transcriptase (RT) enzyme. Upon entry into the cell, the reverse transcriptase is produced from the viral RNA genome using the host cellular machinery. The RT then copies the single-stranded RNA genome into a single-stranded DNA, which then produces a double-stranded viral DNA genome. The double-stranded viral DNA genome is referred to as the **provirus**, which gets incorporated into the host genome from where it keeps producing more retrovirus particles with single-stranded RNA genomes.

1.3 DNA DOUBLE HELIX

The structure of the DNA double helix and its building blocks are described in all biology textbooks. Here, some other aspects are also highlighted, including the information in Box 1.1. DNA is a **double-stranded right-handed** helix; the two strands are **complementary** because of complementary base pairing, and **antiparallel** because the two strands have opposite 5′−3′ orientation (Figure 1.1A). The diameter of the helical DNA molecule is 20 Å (=2 nm). The helical conformation of DNA creates the alternate **major groove** and **minor groove** (Figure 1.1B).

1.3.1 Structural Units of DNA

DNA is composed of structural units called **nucleotides** (deoxyribonucleotides). Each nucleotide is composed of a pentose sugar (2′-deoxy-D-ribose); one of the four nitrogenous bases—adenine (A), thymine (T), guanine (G), or cytosine (C); and a phosphate. The pentose sugar has five carbon atoms and they are numbered 1′ (1-prime) through 5′ (5-prime). The base is attached to the 1′ carbon atom of the sugar, and the phosphate is attached to the 5′ carbon atom (Figure 1.1A). The sugar and base form a **nucleoside**, whereas nucleoside plus phosphate makes a nucleotide. Hence, nucleoside = sugar + base, whereas nucleotide = sugar + base + phosphate. Table 1.1 shows the naming of nucleosides and nucleotides.

BOX 1.1

1. The major grooves in DNA can bind proteins. This is an important property of DNA structure because the major grooves in the upstream regulatory regions of a gene bind transcription-regulatory proteins. For example, for Zn-finger transcription factors, each Zn finger recognizes and binds to a specific trinucleotide sequence in the major groove of DNA.[2]
2. Any double-stranded nucleic acid (whether DNA double strand, DNA−RNA hybrid double strand, or RNA−RNA double strand) is antiparallel in nature. The complementary and antiparallel nature of double-stranded nucleic acids is an important property to remember while designing synthetic oligonucleotides for hybridization (probes or primers).
3. By convention, nucleic acid (DNA or RNA) sequence is written 5′→3′ from left to right, such as 5′-ATGTAAGCAC-3′. If the 5′→3′ designation is not mentioned, it is assumed that the sequence has been written in a 5′→3′ direction, following convention.

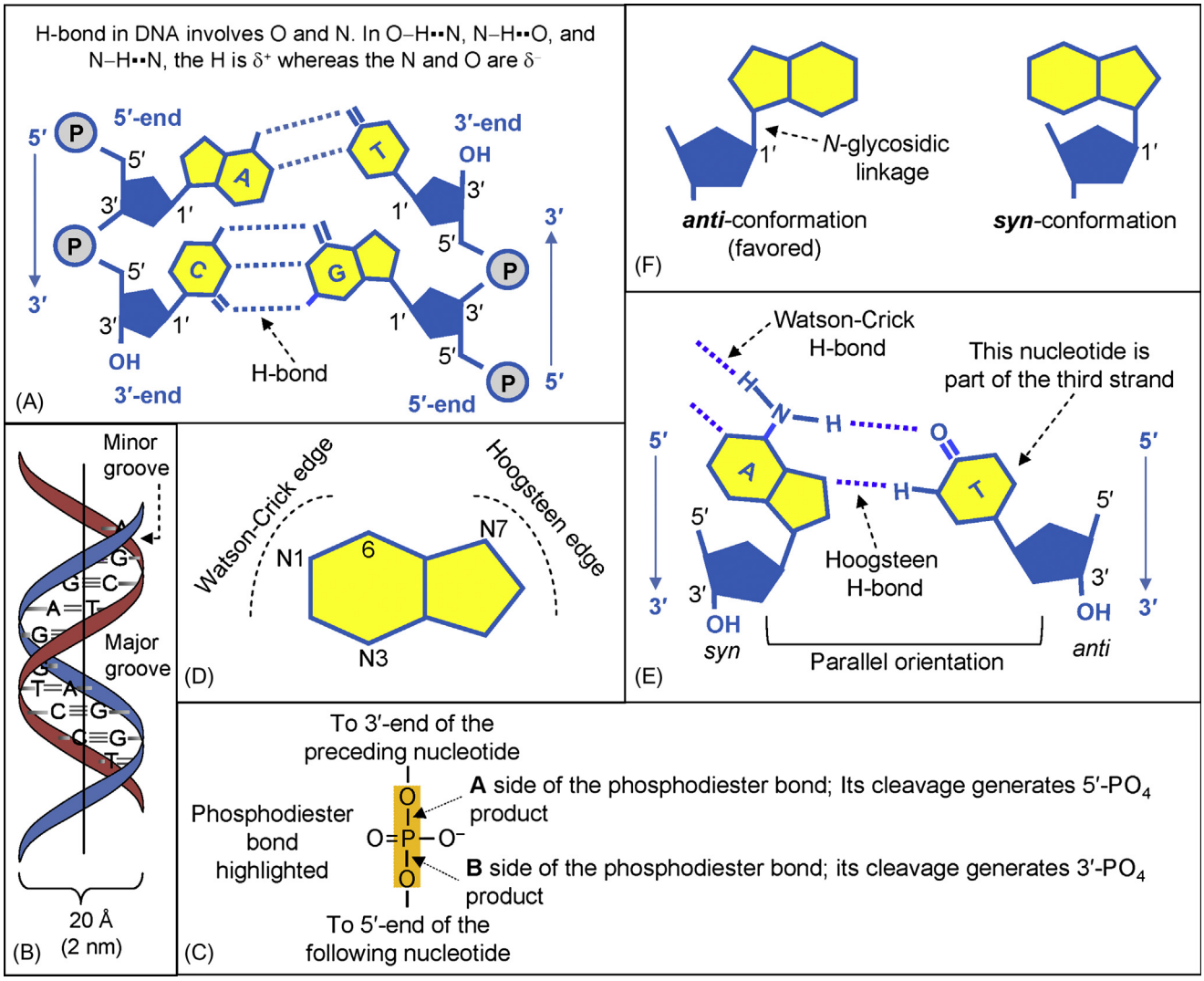

FIGURE 1.1 **DNA structure.** (A) Two nucleotides of the DNA double helix, showing their antiparallel orientation, two H-bonds between A and T and three H-bonds between G and C; (B) the DNA double helix showing the major and minor grooves as well as the diameter of the molecule; (C) the convention of classifying the two sides of the phosphodiester bond and the products generated from their cleavage; (D) the front side (Watson–Crick edge) and the back side (Hoogsteen edge) of a purine; (E) how Hoogsteen H-bonding aids in the formation of the triple helix (see Section 1.3.3); (F) the *anti* and the *syn* conformations of bases around the *N*-glycosidic bond.

TABLE 1.1 Naming of Nucleosides and Nucleotides

Base	Nucleoside (base + sugar)	Nucleotide (base + sugar + phosphate)
Adenine	Deoxyadenosine (sugar = deoxyribose)	Deoxyadenylic acid OR deoxyadenosine monophosphate
Guanine	Deoxyguanosine (sugar = deoxyribose)	Deoxyguanylic acid OR deoxyguanosine monophosphate
Cytosine	Deoxycytidine (sugar = deoxyribose)	Deoxycytidylic acid OR deoxycytidine monophosphate
Thymine	Deoxythymidine (sugar = deoxyribose)	Deoxythymidylic acid OR deoxythymidine monophosphate
Uracil (in RNA)	Uridine (in RNA) (sugar = ribose)	Uridylic acid OR uridine monophosphate

Each nucleotide in DNA (as well as in RNA) has one replaceable hydrogen, which is what makes the DNA (and RNA) acidic.

1.3.2 Linkage between Nucleotides

The nucleotides are joined by 5′–3′ phosphodiester linkage; that is, the 5′-phosphate of a nucleotide is linked to the 3′-OH of the preceding nucleotide by a phosphodiester linkage. In a linear DNA molecule, the 5′-end has a free phosphate and the 3′-end has a free OH group (Figure 1.1A). Each phosphodiester bond has two sides: a 3′-side that is linked to the 3′-end of the preceding nucleotide, and a 5′-side that is linked to 5′-end of the following nucleotide. The 3′-side is called

the **A side** by convention and its cleavage generates a 5′-PO_4 product. The 5′-side is called the **B side** by convention and its cleavage generates a 3′-PO_4 product (Figure 1.1C).

1.3.3 Base-Pairing Rules, Double Helix, and Triple Helix

In the double-stranded DNA, A pairs with T by two hydrogen bonds and G pairs with C by three hydrogen bonds (Figures 1.1A and 1.1B); thus GC-rich regions of DNA have more hydrogen bonds and consequently are more resistant to thermal denaturation. Each **nucleotide pair** (A—T and G—C) has a molecular weight of approximately 660 Da (sodium salt; 610 without sodium). In the helical double-stranded DNA molecule, the sugar—phosphate backbone lies outside and the bases are inside. Base pairs are stacked and horizontal; hence they are perpendicular to the axis of DNA. Because of the stacked nature of the base pairs in DNA, spatially flat molecules can intercalate between them. Of the four bases, A and G are **purines** whereas T and C are **pyrimidines**. In double-stranded DNA, a purine pairs with a pyrimidine (A with T and G with C). Therefore, total amount of purine should equal total amount of pyrimidine; in other words, the purine/pyrimidine ratio should be 1.0 or close to 1.0. This purine—pyrimidine equivalence in double-stranded DNA is called **Chargaff's rule**.

In the bases, the side with the N1 position of the heterocyclic ring is the "front," also called the **Watson—Crick edge** (Figure 1.1D); the opposite side is the "back," also called the **Hoogsteen edge**. Purines have an imidazole ring, which forms the "back"; so in purines, the N7 position of the imidazole ring is part of the Hoogsteen edge (Figure 1.1D). The Hoogsteen edge of the bases is located towards the edge (outside) of the DNA double helix, whereas the Watson—Crick edge is internal. In normal base pairing in DNA and RNA (Watson—Crick base pairing), the Watson—Crick edge (i.e. the front) of the two complementary bases is involved. However, the Hoogsteen edge provides an additional hydrogen bonding site. Therefore, the A—T and G—C base pairs in the normal double helix can form additional hydrogen bonds (**Hoogsteen hydrogen bonds**) to give rise to a triple helix involving the Hoogsteen edge of the purines, i.e. N7 of A and G for the third strand (Figure 1.1E). Hoogsteen hydrogen bonds can also form in RNA. In nucleic acids, the presence of a stretch of homopurine allows a stretch of homopyrimidine to hybridize through Hoogsteen hydrogen bonding to form a section of **DNA triple helix**. *The homopyrimidine-containing third strand is oriented parallel to the oligopurine strand* (Figure 1.1E), *whereas the homopurine-containing third strand is oriented antiparallel to the oligopurine strand* (see Box 1.2).[3—5]

For bases, two conformational variations are possible. The bond joining the 1′-carbon of the deoxyribose sugar to the base is the ***N*-glycosidic bond**. Rotation about this base-to-sugar glycosidic bond gives rise to ***syn*** and ***anti*** conformations. The *anti* conformation is the most common one (Figure 1.1F); however, the *syn* conformation can trigger the formation of triple helix (Figure 1.1E) and also play a role in transversion mutation (see Molecular basis of mutation, Section 2.3.1 in Chapter 2).

1.3.4 Single-Stranded DNA

Many DNA viruses have single-stranded DNA (for example, φX-174, parvoviruses). RNA viruses have RNA as the genetic material, and the RNA genome can be single or double stranded. Single-stranded DNA does not have base equivalence and hence does not follow Chargaff's base equivalence rule.

BOX 1.2

1. Each phosphate has three replaceable H^+; phosphodiester-bond formation between two nucleotides leaves one replaceable H^+. These replaceable H^+ make the DNA (and RNA) acidic (Figures 1.1 and 1.3).
2. The intercalation property of spatially flat molecules is utilized to visualize DNA (and RNA) in a gel using flat aromatic molecules that fluoresce under UV, such as ethidium bromide and acridine orange. The intercalation of these molecules can also cause frameshift mutation during DNA replication.
3. The purine—pyrimidine equivalence can be utilized to determine if a DNA molecule from an unknown source is double stranded or single stranded. In a double-stranded DNA molecule, the purine/pyrimidine ratio should be 1.0 (or close to 1.0); in contrast, in a single-stranded DNA molecule this equivalence is lacking.
4. The differential thermal stability of AT-rich versus GC-rich regions in double-stranded nucleic acids is taken into consideration while designing oligonucleotides for hybridization for different

BOX 1.2 (*cont'd*)

purposes, such as high-stringency hybridization, primers for polymerase chain reaction (PCR), or for sequencing. For example, an oligoprobe that will be used for high-stringency hybridization can have ≥ 55% G + C content.

5. If the molecular weight of an unknown double-stranded DNA is determined, the total base-pair content of the DNA can be calculated based on the fact that each **nucleotide pair** has an approximate molecular weight of 660 Da. By the same token, if the total number of base pairs in a DNA molecule is known, its molecular weight can be determined as well.
6. Hoogsteen hydrogen bonding can create short transient stretches of triple helix in vivo; triple helix formation can also be induced under experimental conditions. Synthetic oligodeoxynucleotides that can form triple helix have been used in vitro to inhibit gene expression in cells. Triple-helix-forming oligonucleotides coupled to DNA-modifying agents can be introduced into cells to modify the DNA target in a highly sequence-specific manner. This tool can be used to introduce genome modification, modulate specific gene expression, or even repair DNA.[6,7]

1.3.5 Base Sequence and the Genetic Code

The genetic information—that is, the genetic code with information for the amino acid sequence of the protein—lies in the sequence of bases in DNA. Genetic code exists in the form of a sequence of three bases; each three-base sequence is called a **codon**, which codes for an amino acid. Transcription of mRNA copies the codons from DNA to mRNA, which is translated to yield the protein (polypeptide) product. ATG in DNA (corresponding to AUG in RNA) is the start codon that codes for methionine. Translation begins by recognizing the start codon and incorporating methionine as the first amino acid. Similarly, TAG (**amber**), TGA (**opal**), and TAA (**ochre**) (corresponding to UAG, UGA, and UAA, respectively, in mRNA) are the three stop codons that do not code for any amino acids (exceptions to this rule are discussed below). In addition to being triplet (read as three-nucleotide codons), genetic code is (almost) **universal**, **non-overlapping** (adjacent codons do not share nucleotides), and **degenerate** (most amino acids can be coded by more than one codon). There are 64 (4^3) possible codons (61 coding and 3 noncoding). Genetic code normally codes for 20 standard amino acids. The two known cases of direct incorporation of non-standard amino acids are that of **selenocysteine** (the 21st amino acid) and **pyrrolysine** (22nd amino acid). Selenocysteine has been found in lower as well as higher organisms, including mammals, while pyrrolysine has so far been found in certain archaebacteria. Both these amino acids are encoded by stop codons; selenocysteine is encoded by UGA and pyrrolysine is encoded by UAG in mRNA.

1.4 CONFORMATIONS OF DNA

There are three major conformations of DNA: **B-DNA**, **A-DNA**, and **Z-DNA**. The DNA structure that Watson and Crick proposed was the B form of DNA (B-DNA), and this is the physiological form of DNA. In B-DNA, the diameter of the helix is 2 nm (=20 Å). Each pitch—that is, one complete turn (360°)—is 3.4 nm (=34 Å) long and contains 10 base pairs. A-DNA has been identified in vitro under different salt concentrations, as well as in DNA−RNA hybrids. It is also a right-handed helix. The diameter of the helix is 2.3 nm (=23 Å). Each pitch is 2.6 nm (=26 Å) and contains 11 base pairs. So, for a given length, the A-form is wider and shorter than the B-form. Z-DNA is a **left-handed helix** (Z = zigzag). This form has been identified both in vitro and within the cell. Small, localized regions within the physiological B-form of DNA can attain a left-handed conformation. Formation of the left-handed Z-DNA conformation is dictated by regions of alternating purines and pyrimidines residues, such as 5′-GCGCGCGCGCGCGCGC-3′. In Z-DNA, the diameter of the helix is 1.8 nm (=18 Å). Each pitch is 3.7 nm (=37 Å) long and contains 12 base pairs. Thus, the Z-form is narrower and longer than the B-form. It is thought that local Z-DNA conformations may play important roles in gene transcription.

1.5 TYPICAL EUKARYOTIC GENE STRUCTURE

According to the classical view of transcription, for any given gene, one of the two strands of DNA is

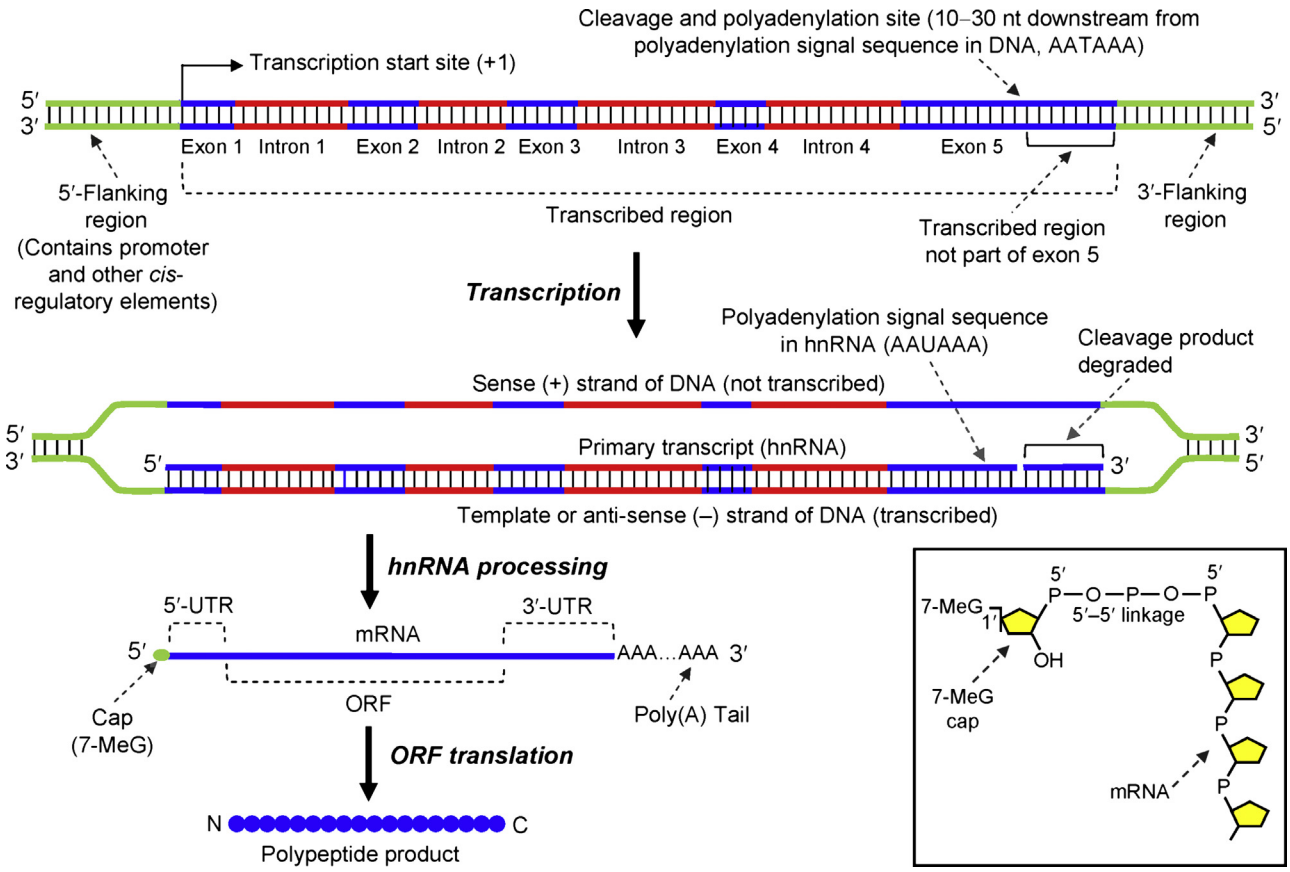

FIGURE 1.2 **Gene–hnRNA–mRNA–protein relationship.** Exon 1 is noncoding. Thus, the 5′-untranslated region (5′-UTR) is derived from exon 1, and the 3′-UTR is derived from the noncoding part of exon 5, which is the last and the longest exon. The sense strand of DNA has a "T" where the mRNA has "U"—for example, the poly(A) signal sequence in the sense strand is AATAAA, but in RNA it is AAUAAA. The transcription initiation site is +1 and the base to the left (upstream) of it is −1; there is no 0 position. Also, note that RNA polymerase transcribes well beyond the poly(A) site; this extra part of the transcript is degraded and does not form part of the last exon. Inset shows the mRNA cap (7-MeG) and its 5′–5′ linkage with the first base of mRNA. nt, nucleotide; ORF, open reading frame.

transcribed, the other is not[a]. The DNA strand that is NOT transcribed is called the **sense** or **plus** (+) or **coding** strand because it has the same sequence as that of the mRNA (except for U in RNA and T in DNA)—that is, the same sequence of codons in the same 5′→3′ direction, so that the polypeptide sequence can be predicted from the sense strand sequence (see Box 1.3). In contrast, the strand that is transcribed is called the **template** or **antisense** or **minus** (−) or **noncoding** strand because its sequence is complementary to the coding sequence; hence, the polypeptide sequence cannot be predicted from the template strand sequence. A typical mRNA-coding eukaryotic gene has three major parts: a transcribed region, a 5′-flanking region, and a 3′-flankng region (Figure 1.2). In eukaryotes, different types of RNAs are transcribed from the DNA by different RNA polymerases: RNA polymerase I (pol I) transcribes ribosomal RNA (rRNA), RNA polymerase II (pol II) transcribes messenger RNA (mRNA), RNA polymerase III (pol III) transcribes transfer RNA (tRNA). For mRNA, the primary transcript that contains both exons and introns is called the **heterogeneous nuclear RNA** (**hnRNA**) or **pre-mRNA**. The hnRNA is processed to remove the introns (**splicing**), add a 7-methyl guanine **cap** at the 5′-end by 5′–5′ linkage (Figure 1.2 inset), and add a **poly(A) tail** at the 3′-end, which is about 200 bp long in mammals.

[a]The classical view of transcription is an oversimplification. Deep sequencing and global transcriptome analysis have demonstrated that a significant proportion of the genome can produce both sense and antisense transcripts. When the sense and antisense transcripts are produced from the opposite strands of DNA in the same genomic locus, the antisense transcript is called a ***cis*-antisense** transcript because its target is the sense transcript. In contrast, ***trans*-antisense** transcripts are transcribed from a different location than their targets (e.g. microRNAs).

1.5.1 Transcribed Region

The nucleotide sequence of a gene that is transcribed into mRNA is composed of discrete sequences called **exons** and **introns**. Introns are also known as intervening sequences (abbreviated as **IS**) (Figure 1.2). After transcription of the gene, a longer primary transcript (the hnRNA or pre-mRNA) is produced. The hnRNA has the same exon−intron organization as the gene: exons are interrupted by introns. The hnRNA is processed to produce the mature mRNA. Exons are maintained in the mature mRNA, while introns are spliced out (in most cases). The structural unit of mRNA is the ribonucleotide (Figure 1.3). Introns do not contain information for the coding of the polypeptide. However, some introns, usually at the 5′-end of the gene, contain signals for transcriptional regulation. Introns of many genes also contain **nested genes** that have distinct expression profiles.[8] In mRNAs, a few terminal exons are noncoding, whereas the internal exons code for amino acids. These terminal noncoding exons form the 5′- and 3′-untranslated regions (UTRs) of the mRNA. In most mRNAs, the last exon (at the 3′-end) is usually the longest of all exons, and is partially coding (see Box 1.4).

1.5.1.1 Intron-Splicing Signals

Most introns in genes have GT at the 5′-splice site (in the DNA sense strand; hence GU in the hnRNA), called the **splice donor** site, and AG at the 3′-splice site, called the **splice acceptor** site. These introns are referred to as GT−AG introns. However, introns may also contain GC or AT as the splice donor sites, and AC as the splice acceptor site (hence, GC−AG introns, AT−AC introns).

In most eukaryotic genes, the nucleotides surrounding the splice donor and acceptor sites show a great degree of conservation. The usual nucleotide distribution around the splice sites is as follows:

5′-splice site: 5′-. . .NNNAG**gt**annn. . .3′ (**gt** = splice donor site in the intron; N = any nucleotide in the exon; n = any nucleotide in the intron; bases underlined are usually conserved; AG are the last two bases of the preceding exon, and a is the base that immediately follows the splice donor site).

3′-splice site: 5′-. . .nnnc**ag**NNN. . .3′ (**ag** = splice acceptor site in the intron; N = any nucleotide in the following exon; n = any nucleotide in the intron; the base underlined is usually conserved; c is the base immediately preceding the splice acceptor site).

Two other important sequence elements are the **branch point** and the **polypyrimidine tract** in the introns. The branch point is located 20−50 nucleotides upstream from the splice acceptor site. The consensus sequence of the branch point site is (C/T)(T/C)(A/G)**A**(C/T), in which the **A**-residue is conserved in all genes. This **A**-residue is called the branch point and it plays a crucial role in splicing. The polypyrimidine tract is located downstream from the branch point.

BOX 1.3

1. An easy way to remember the sense and antisense designations is to remember just one fact: that the sequence of mRNA is sense. This is because the codons can be found in the coding sequence of mRNA; as a result the amino acid sequence of the polypeptide can be predicted from the mRNA coding sequence. Hence, any sequence that is same as the mRNA sequence along with the same 5′→3′ polarity is also sense. That is why the DNA strand that has the same sequence and polarity as the mRNA is also sense. Likewise, any sequence that is complementary to the mRNA sequence, along with the opposite 5′→3′ polarity, is antisense. Hence, the template DNA strand is antisense (Figure 1.4A).
2. By the same token, the probe used to detect mRNA in northern blot or in situ hybridization is antisense because it is complementary and has an opposite polarity to the mRNA. When designing antisense DNA oligoprobes for RNA or DNA hybridization, the complementary and antiparallel sequence of the sense strand of DNA is used. For example, in Figure 1.4, the mRNA partial sequence shown is 5′-AUG UGU AGA UCG AUG A-3′. That region of the antisense DNA probe will have the sequence 3′-TAC ACA TCT AGC TAC T-5′. Following convention, the DNA probe sequence has to be rewritten in a 5′→3′ direction from left to right. Hence, this DNA probe partial sequence will be rewritten (for reporting the sequence) as 5′-TCA TCG ATC TAC ACA T-3′ (Figure 1.4B).
3. In the nucleotide databases, such as in National Center for Biotechnology Information (NCBI), DNA Data Bank of Japan (DDBJ), or The European Molecular Biology Laboratory (EMBL), the reported mRNA sequences do not contain U but instead contain T. This is because the mRNA sequence is reported as the sense strand of the cloned complementary DNA (cDNA).

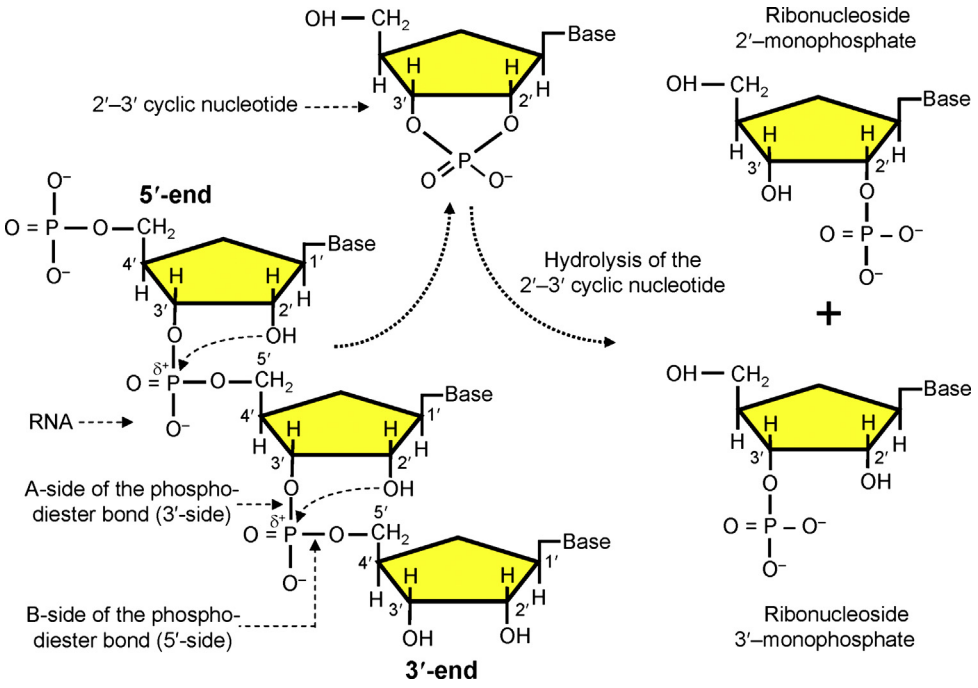

FIGURE 1.3 **Alkaline hydrolysis of RNA.** In an alkaline pH, the OH^- can abstract the H from the 2′-OH of ribose, generating the nucleophile 2′—O^-, which carries out a nucleophilic attack on the δ^+ P of the phosphate. This results in the cleavage of the phosphodiester bond and the formation of 2′−3′ cyclic nucleotide; the cyclic nucleotide hydrolyzes into ribonucleoside 2′- and 3′-monophosphate end products.

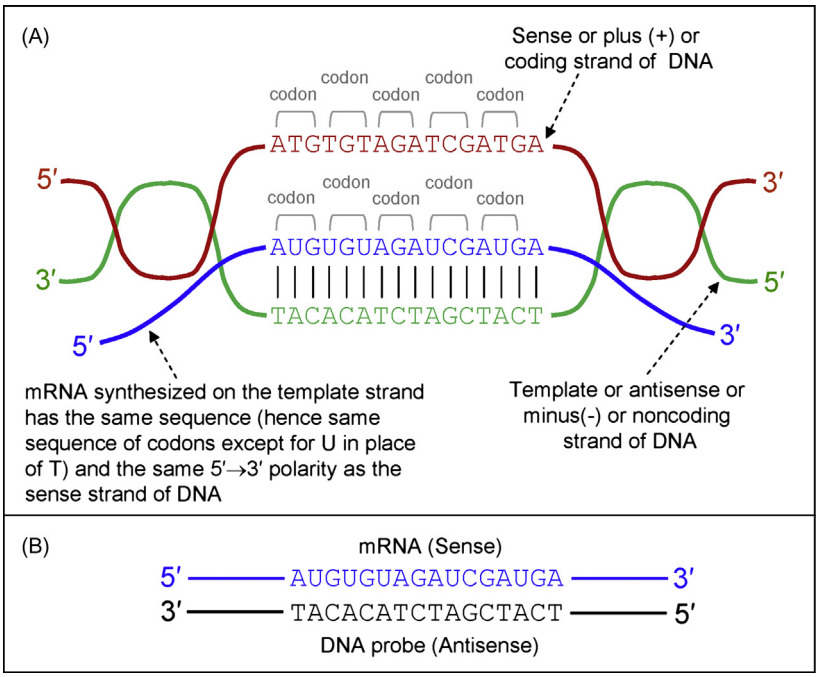

FIGURE 1.4 **Sense and antisense strands of DNA.** (A) The two strands of DNA have been drawn in different colors so that their respective 5′- and 3′-ends could be easily distinguished. The figure shows that mRNA and the sense strand have the same sequence (except for "U" in RNA and "T" in DNA) and the same 5′—3′ polarity. (B) The mRNA and antisense probe relationship.

BOX 1.4

1. Sometimes an intron may be retained in the mature mRNA and perform specific regulatory functions. For example, migration stimulatory factor (MSF) is a truncated oncofetal isoform of fibronectin. Two types of MSF mRNAs have been detected: a shorter 2.1-kb[b] transcript and a longer 5.9-kb transcript, which differ only in the length of their 3′-UTRs. In the smaller transcript, the intron-derived 30-nucleotide (nt) coding sequence is followed by a 165-nt intron-derived 3′-UTR. This makes a total of 195-nt intron-derived sequence in the smaller transcript.[9] This intron-derived 3′-UTR also provides the polyadenylation signal. The smaller transcript is transported to the cytoplasm and eventually secreted, while the larger transcript is retained in the nucleus.
2. After a gene is cloned and sequenced, the exon–intron boundaries are identified by comparing the gene sequence with its complementary DNA (cDNA) (mRNA) sequence. Identification of the exon–intron boundaries of a gene is essential when attempting to manipulate the DNA, such as making a gene-targeting construct.
3. The majority of internal exons in vertebrate genes are less than 300 bp; the average length being 135 bp; exons larger than 800 bp are rare.[10]
4. For most genes, the last exon (at the 3′-end) is the longest exon (could be well over 1 kb) and partially coding.
5. For most genes, the 5′-UTR is derived from more than one exon. Of these 5′ noncoding exons, the most downstream one is usually partially noncoding because the open reading frame (ORF) begins at some place in this exon, making it partially noncoding and partially coding.
6. For most genes, the 3′-UTR is three to five times longer than the 5′-UTR, particularly in vertebrates.
7. In vertebrates, exons are small and introns are large. In contrast, in lower eukaryotes, the opposite is true.[11]
8. The transcription start site (+1) in most genes begins with a purine (mostly an "A").

[b]kb, kilobase = 1000 bases; Mb, megabase = 1000 kb; Gb, gigabase = 1000 Mb. In the context of DNA, these mean base pairs (hence, kbp, Mbp, and Gbp).

BOX 1.5

Knowledge of the intron phases helps predict which exon(s) can or cannot be targeted for alternative splicing. Exceptions to this rule have also been reported in the literature. For example, the alternative splicing of *rat liver-specific organic anion transporter* pre-mRNA, generating a functional mRNA, involves the removal of exon 10, which is an asymmetrical exon flanked by a phase 1 and a phase 2 intron. The creation of a frameshift mutation in this unusually spliced mRNA is averted by retaining 91 bp from the 5′-end of exon 10 in the mature mRNA.[12]

1.5.1.2 Effect of Intron Phase on Alternative Splicing

Introns can be divided into three types based on phases: **phase 0**, **phase 1**, and **phase 2**. A phase 0 intron does not disrupt a codon, a phase 1 intron disrupts a codon between the first and second bases, whereas a phase 2 intron disrupts a codon between the second and third bases. An exon flanked by two introns of the same phase is called a **symmetrical exon**, whereas an exon flanked by two introns of different phases is called an **asymmetrical exon**. Intron phase determines which exons may or may not be targeted for alternative splicing. With a few rare exceptions, exons that are subjected to alternative splicing are always symmetrical exons—that is, exons flanked by same-phase introns. In contrast, asymmetrical exons—that is, exons flanked by different-phase introns—cannot be alternatively spliced because such alternative splicing will throw the normal open reading frame (ORF) out of frame beyond the 3′-splice site (Figure 1.5). Such frameshift results in the creation of premature stop codon and truncation of the ORF. Intron phase determines exon shuffling potential, which determines protein domain shuffling during protein evolution and the evolution of organismal complexity (discussed in Chapter 2; see Box 1.5).

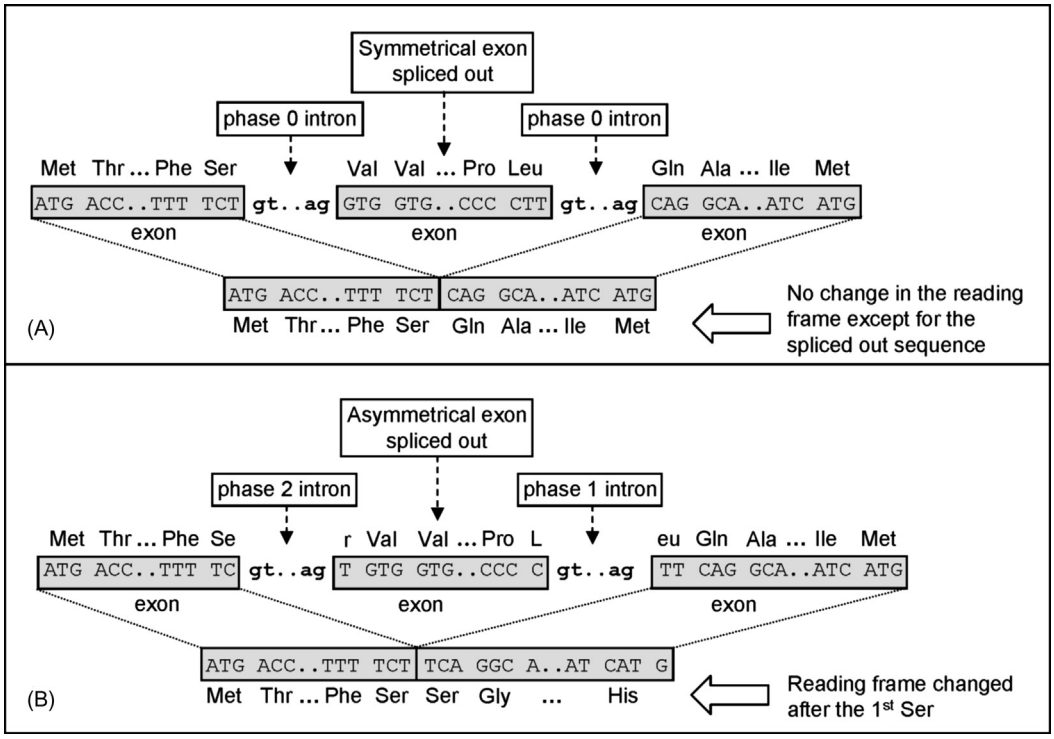

FIGURE 1.5 **The effect of intron phase on alternative splicing.** (A) Alternative splicing involving the removal of a symmetrical exon (flanked by introns of the same phase; 0–0) does not cause a frameshift in the ORF except for the deletion of the amino acids encoded by the removed exon; (B) alternative splicing involving the removal of an asymmetrical exon (flanked by introns of different phase; 2–1) causes a frameshift in the ORF downstream from the 3′-splice site. Such frameshift results in the creation of a premature stop codon and truncation of the ORF.

1.5.1.3 Evolution of Introns

After the initial discovery of introns in 1977, the **introns-early theory** was proposed to explain the origin and evolution of introns. According to the introns-early theory, introns were present as intergenic regions in the genome of the common ancestor of prokaryotes and eukaryotes. These intergenic genomic regions were subsequently lost in all prokaryote lineages; in contrast they were maintained in eukaryotes as introns owing to the appearance of the spliceosomal machinery. Walter Gilbert suggested that the presence of introns allowed exon shuffling, which resulted in genomes being more complex and diversified. The accumulation of genomic data has helped reconstruct the evolutionary history of introns and replace the introns early theory with the **introns-late theory**. According to the introns-late theory, self-splicing introns (also known as **retrointrons**) first invaded eukaryotic genomes, and spliceosomal introns were subsequently derived from self-splicing introns. Hence, spliceosomal introns only appeared in eukaryotes. Spliceosomal machinery evolved as a means of removing spliceosomal introns. Therefore, the last common ancestor of eukaryotes had a spliceosomal-intron-rich genome. The intron-containing genomes probably spread due to **population bottlenecks**[c]. Further massive intron invasion of the genome was likely limited only to those genomes that underwent significant evolutionary innovations. Intron loss in many lineages also occurred, resulting in the present-day intron-poor species.[13,14]

Introns-late theory envisages that early introns had no functions; hence their presence was deleterious for the genomes. However, early introns were transcribed and were free from selective constraints; hence, at some point during evolution, they might have gained

[c]Population bottleneck is a phenomenon in which the population size is drastically reduced through events like environmental disaster, habitat destruction, or massive predation and hunting. As a result, only a small fraction of the genetic diversity of the original population survives. When the population multiplies, the surviving genetic diversity spreads in the population. Thus, if the intron-containing genome survived through a population bottleneck, it subsequently spread in the resulting population. In general, population bottleneck results in a drastic reduction of the gene pool and genetic diversity in the resulting population. Owing to the loss of genetic variation, the new population could be genetically distinct from the original population. Loss of genetic diversity, particularly in a small population, can cause genetic drift and rare alleles face increased chance of being lost.

some functions. One of the best known functions of introns is their ability to increase transcription and ultimately protein expression of intron-bearing genes compared to intronless genes. In making transgenic organisms, particularly transgenic plants, specific introns are frequently included in the construct to increase the expression of the transgene.

Introns are now known to mediate their function by modulating every possible step of transcription: initiation, elongation, termination, mRNA maturation, nuclear export, and mRNA stabilization. The mechanism of action of many introns is not known. However, the functions of introns can be sequence-dependent, length-dependent, position-dependent, and splicing-dependent.[15]

1.5.2 5′-Flanking Region of Transcribed Genes

The 5′-flanking region of transcribed genes contains the **promoter**. The promoter contains specific sequences for binding the proteins necessary for transcription by RNA polymerase. The specific sequence in the promoter that positions the pol II is called the **TATA box** (consensus 5′-TATAAA-3′; some variants exist). Typically, the TATA box is located 25−30 bp upstream of the transcription start site (that is, −25 to −30 bp position), and for any given gene the position of the TATA box is fixed. However, many gene promoters lack the TATA box (TATA-less promoters). Accurate positioning of pol II in TATA-less promoters is thought to be mediated by two other *cis*-acting sequence elements, the **initiator** element (**Inr**) and the **downstream promoter element** (**DPE**). Inr has a consensus sequence of Y-+1-N-T/A-Y-Y (where Y is a pyrimidine, +1 is the transcription initiation site, N is any nucleotide), and DPE has a consensus sequence of $(A/G)_{+28}G(A/T)(C/T)(G/A/C)_{+32}$. Therefore, Inr occurs around the transcription start site and DPE occurs between 28 and 32 bases downstream from the transcription start site. Many variants of the Inr sequence have been reported. DPE has been most extensively studied in *Drosophila*. Some other sequences in the promoter that are found in most genes are the CAAT-box (around −75 to −80 bp position) and the GC-box (around −90 bp position).

Various regions of the promoter have been termed the core (or basal), proximal, and distal promotor depending on their distance from the transcription start site. The **core promoter** is about 35 bp long and extends 35 bp upstream or downstream from the transcription site (−35 to +35), the **proximal promoter** is around 250 bp long, whereas the **distal promoter** is located further upstream. Therefore, the TATA box, Inr, and DPE are all contained within the core promoter, whereas the CAAT-box and the GC-box are contained within the proximal promoter. Core, proximal, and distal promoter elements cooperate to regulate transcription.

The proximal promoter contains additional *cis*-acting sequences that are necessary for the regulation of gene expression in response to specific stimuli. These sequences are called **response elements** or **regulatory elements** (**RE**). For example, genes that are induced by glucocorticoids have a glucocorticoid response element (GRE) in their promoters. Many such response elements have been identified so far in a number of animal and plant gene promoters. These response elements bind specific transcription regulatory proteins called transcription factors that control gene expression. Regulatory elements can also be found far upstream of the TATA box, far downstream in the 3′-flanking sequence, and even within introns. These elements typically act as enhancers because they significantly upregulate the expression of genes (see Box 1.6).

1.5.3 3′-Flanking Region of Transcribed Genes

Although it is often said that the 3′-flanking region contains the transcription termination signal, eukaryotic pol II does not terminate transcription at any definitive

BOX 1.6

Promoter-bashing experiments help identify the importance of specific promoter sequences in regulating gene expression. These experiments make use of deletion mutations to narrow down the region of interest; then individual bases are mutated to define the core functional sequence involved in regulating transcription. Bioinformatic software uses the available information on various identified transcriptional activator- or repressor-binding sequences, and scans the 5′-flanking sequences of a gene to predict putative binding sites in the promoter. However, many of the putative binding sites predicted through bioinformatic analysis may turn out to have no effect on transcription when verified through promoter-bashing experiments. Thus, predicted regulatory sequences are only a rough guide and need functional verification through experimentation.

termination signals in the DNA. For most eukaryotic protein-coding genes, pol II transcribes the template strand 500−2000 nucleotides beyond the polyadenylation site (Figure 1.2). Transcription termination is facilitated by a number of protein factors (such as Cleavage and Polyadenylation Specificity Factor (CPSF), Cleavage Stimulation Factor (CStF), etc.) that become associated with the pol II as soon as the enzyme leaves the promoter. These factors, along with capping and splicing factors, ride on the C-terminal domain (CTD) tail of pol II. Transcription of the poly(A) signal sequence triggers the endonucleolytic cleavage of the nascent transcript, degradation of the downstream cleavage product, and termination of transcription. The pausing of pol II downstream from the poly(A) site appears to be an obligatory step leading to termination, which involves the displacement of pol II from the template. *The 5′- and 3′-ends of a gene are the same as the 5′- and 3′-ends of the sense strand*.

1.6 MUTATIONS IN THE DNA SEQUENCE

The sequence that codes for a polypeptide is referred to as the coding region or **open reading frame** (**ORF**). Various mutations in the ORF may or may not lead to changes in the amino acid sequence in the polypeptide product. If a mutation in DNA leads to an amino acid change in the polypeptide, it is called a **missense** or **nonsynonymous** mutation; if a mutation does not lead to an amino acid change in the polypeptide, it is called a **silent** or **synonymous** mutation. Traditional wisdom assumes that a synonymous mutation does not alter the protein function because there is no change in the amino acid. However, recent findings indicate that, in many proteins, synonymous mutations may also alter protein function because they result in an altered conformation of the protein. Because protein folding is a co-translational process, proper protein folding is tightly linked to the speed of translation. Synonymous mutations that affect codon usage may disrupt this process resulting in a wrongly folded polypeptide. In fact, some human diseases could be linked to such synonymous mutations.[16]

1.7 SOME FEATURES OF RNA

In traditional molecular biology, a discussion on RNA focused on three types of RNA associated with protein synthesis: ribosomal RNA (rRNA), messenger RNA (mRNA), and transfer RNA (tRNA), of which rRNA and tRNA are noncoding, whereas mRNA is protein coding. The world of functional noncoding RNA molecules has since been greatly expanded (discussed later). As mentioned above, RNA is the genetic material in retroviruses. An RNA molecule is single stranded, except in regions where base complementarity makes the molecule fold back on itself forming double-stranded segments. Like DNA, RNA is also composed of nucleotides (ribonucleotides). However, there are two differences from DNA: the sugar is ribose and the base uracil ("U") is present instead of "T"; thus the base pairing is between "A" and "U." Of the three RNAs associated with translation (rRNA, mRNA, and tRNA), the following discussion focuses on mRNA.

1.7.1 Instability of mRNA

Apart from the ubiquitous presence of the enzyme RNAse that can easily degrade mRNA, the structure of mRNA itself also contributes to its instability. The ribose sugar makes RNA less stable than DNA, especially at alkaline pH. At alkaline pH, the 2′-OH of the ribose sugar undergoes alkaline hydrolysis, which results in the breakage of the phosphate bond between adjacent nucleotides, and formation of the 2′−3′ cyclic nucleotide (Figure 1.3). Hydrolysis of this 2′−3′ cyclic nucleotide gives rise to a mixture of ribonucleoside 2′- and 3′-monophosphate products. In contrast, in DNA the 2′ carbon has an H instead of an OH, which prevents the formation of the 2′−3′ cyclic nucleotide; this prevents alkaline hydrolysis and makes DNA stable at alkaline pH. At acidic pH, however, phosphodiester bond hydrolysis occurs in both DNA and RNA. Because RNA undergoes rapid alkaline hydrolysis, particularly around 37°C, use of NaOH (even ice-cold) to denature RNA is not recommended.

1.7.2 5′- and 3′-Untranslated Regions of mRNA

A typical eukaryotic mRNA has three regions: a 5′-untranslated region (5′-UTR), a coding region or ORF, and a 3′-untranslated region (3′-UTR). The translational start codon is AUG, and there is one of the three translational stop codons, UAA, UGA, and UAG. The 5′-end of mRNA has the cap (7-methyl GTP) attached to the first base through a **5′−5′ linkage**. The 5′- and 3′-UTRs are composed of noncoding exons or noncoding parts of partially coding exons, whereas the ORF is composed of coding exons. The last exon at the 3′-end is usually the longest. The 3′-UTR of mRNAs contains the **poly(A) signal** sequence 5′-AAUAAA-3′, which is located 10−30 nucleotides upstream of the polyadenylation site (see Box 1.7). The poly(A) tail is around 200 bp long in mammals. The cap at the 5′-end and the poly

(A) tail at the 3′-end help in translation and also aid in the stability of the mRNA. If the 3′-UTR of an mRNA contains multiple poly(A) signal sequence, the mRNA may undergo **alternative polyadenylation**, producing transcripts with very different stability. Alternatively polyadenylated mRNAs also differ in the length of their 3′-UTRs; they can be observed in different tissues or at different developmental stages where the half-life of the same mRNA may markedly vary.[17] Many mRNAs with more than one poly(A) signal sequence have been reported in the database, but not all of them have been experimentally tested to confirm the generation of alternatively polyadenylated transcripts.

The 5′-UTR of mRNA controls the initiation of translation. An important sequence relevant for translation initiation and identification of the correct AUG codon (translation start codon) is called the **Kozak sequence**, after its discoverer, Marilyn Kozak. The original Kozak sequence described was 5′-CCRCC**AUG**G-3′ where **AUG** is the translation start codon, and R is a purine. Later on, a shorter yet highly effective version of the Kozak sequence was described as 5′-ACC**AUG**G-3′. Although many mRNAs contain the consensus Kozak sequence or some variant of it, there are many other mRNAs that do not contain any Kozak sequence at all.

The 5′-and 3′-UTRs of mRNAs can also regulate gene expression and mRNA stability by interacting with proteins or nonprotein ligands. For example, the expression of feritin mRNA is regulated by the binding of specific regulatory proteins to its 5′-UTR, whereas the stability of transferrin receptor mRNA is regulated by the binding of specific regulatory proteins to its 3′-UTR. In contrast to protein ligands, in bacteria certain mRNAs can regulate gene expression by binding specific nonprotein ligands. The part of the mRNA that binds to the small molecule and acts as the genetic switch is called a **riboswitch**. Some examples include coenzyme-B12-binding riboswitch, flavin mononucleotide (FMN)-binding riboswitch, thiamine or thiamine pyrophosphate (TPP)-binding riboswitch—all located in the 5′-UTR of the relevant mRNAs.[18]

1.7.3 Secondary Structures in RNA

RNA crystallography has revealed the existence of a rich variety of base pairing, giving rise to a multitude of complex tertiary structural motifs. Leontis and Westhof[19] proposed that the planar edge-to-edge hydrogen-bonding interactions between RNA bases involve one of three distinct edges: the **Watson−Crick edge**, the **Hoogsteen edge**, and the **sugar edge** (which includes the 2′-OH). About 60% of the bases participate in canonical Watson−Crick base pairs. The original geometric nomenclature and classification has been recently revisited by Abu Almakarem et al.,[20] who developed a classification scheme that is predicted to help identify recurrent base triplets (referred to as "base triples" in the publication) that can substitute for each other while conserving RNA three-dimensional structure. Hence, the system has applications in RNA three-dimensional structure prediction and analysis of RNA sequence evolution. Taking into consideration the spatial orientations in which bases can interact, Leontis and Westhof identified 12 basic geometric types with at least two H-bonds connecting the bases. In other words, Leontis and Westhof defined 12 base-pair families. Using the combinatorial enumeration of these 12 base-pair families, Abu Almakarem and coworkers predicted the existence of 108 potential geometric base-triple (triplet) families. Searching representative atomic-resolution RNA three-dimensional structures revealed instances of 68 of the 108 predicted base-triple families. Further model building suggested that some of the remaining 40 families may be unlikely to form for steric reasons.

BOX 1.7

1. Bioinformatic analysis of any sequence that might code for a polypeptide will produce a total of six reading frames: three in sense, three in antisense. Of these, one reading frame is always the longest, providing the legitimate ORF. Some software produces only three sense-frame output.
2. The polyadenylation (poly(A)) signal sequence is highly conserved. The canonical poly(A) signal sequence identified in cloned complementary DNA (cDNA)/gene sequence is AATAAA (AAUAAA in the mRNA). The only other known functional variant of the poly(A) signal sequence is ATTAAA (AUUAAA in the mRNA).

1.8 CODING VERSUS NONCODING RNA

In addition to rRNA and tRNA, a few other classes of ncRNAs have been known for some time, such as snRNA (small nuclear RNA), snoRNA (small nucleolar RNA), gRNA (guide RNA), *Xist* (X inactive-specific transcript) and *Tsix* (an antisense regulator of *Xist*), *H19*, *Air*, and *Kcnq1ot1* (potassium channel Q1 overlapping transcript 1). These ncRNAs are very different in length (e.g. 50–70 nucleotides (nt), such as gRNA, to more than 100 kb, such as *Air* ncRNA), and they serve diverse functions. For example, snRNAs are essential for mRNA splicing, snoRNAs are important in methylation of rRNAs, gRNAs are essential in RNA editing, whereas *Xist*, *Tsix*, *H19*, *Air*, and *Kcnq1ot1* are all involved in the epigenetic regulation of gene and genome expression; for example, *Xist* and *Tsix* are involved in X-chromosome inactivation in mammals whereas *H19*, *Air*, and *Kcnq1ot1* are associated with imprinted loci and genomic imprinting. Since the 1990s, the RNA universe has been producing regular surprises that have enriched our idea about RNA's role in gene regulation, and the breadth of the cellular gene regulatory network itself.

1.8.1 Small Noncoding RNA, Long Noncoding RNA, Competing Endogenous RNA, and Circular RNA

In recent years, a new class of ncRNAs, the **small ncRNAs** (~20–30nt long), has been identified as very powerful regulators of gene expression. Examples include **microRNA** (**miRNA**, abbreviated as **miR**), **small interfering RNA** (**siRNA**), and **Piwi-interacting RNA** (**piRNA**).[21,22]

These small ncRNAs are generated through the processing of double-stranded segments of long precursor RNAs. Accordingly, software has been developed to identify putative genomic sequences that may give rise to small ncRNAs, as well as potential target sequences of these putative ncRNAs. These theoretical predictions have to be experimentally confirmed. An ever-increasing number of studies have implicated miRNAs and siRNAs in human health and disease, ranging from metabolic disorders to diseases of various organ systems, including various forms of cancer. More than 30% of all human genes have been predicted to be miRNA targets. Consequently, a number of freely accessible web-based miRNA databases have been developed that contain both predicted and experimentally verified miRNA sequences. One such database is the miRBase (http://microrna.sanger.ac.uk/), which is one of the most comprehensive miRNA databases. Release 19.0 (August 2012) of the miRBase reports a total of 21,264 identified miRNAs in different species, of which 2214 are identified in humans. Examples of some other miRNA databases are:

miRNAviewer (http://cbio.mskcc.org/mirnaviewer/)
miRWalk (http://www.umm.uni-heidelberg.de/apps/zmf/mirwalk/)
MicroRNA.org (http://www.microrna.org/microrna/home.do)
miRGator (http://genome.ewha.ac.kr/miRGator/).

Long noncoding RNAs (**lncRNAs**) are >200 nucleotides in length and do not code for protein. The lncRNAs are the least understood among the ncRNAs, but evidence suggests that they play important roles in a broad range of biological processes.[23] The *Air*, *Xist*, *Tsix*, and *Kcnq1ot1* RNAs discussed above are all lncRNAs. A good lncRNA database can be accessed at http://www.lncrnadb.org/.[24]

Just as an efficient regulatory network should have multiple control points, the regulation of gene expression by miRNAs is further regulated by other RNAs. Two such recently discovered miRNA-regulatory RNAs are competing endogenous RNA (ceRNA) and the most recently reported circular RNA (circRNA). Functionally, both these RNAs antagonize the effects of miRNA. The discovery of these anti-miR RNA molecules will trigger a reevaluation of the model of the RNA regulatory network, and the gene regulatory potential of miRNAs.

As the name implies, **competing endogenous RNAs** (**ceRNAs**) are noncoding RNA molecules that contain binding sites for miRNAs, referred to as miRNA response elements (MREs), and thus compete with the miRNA targets to bind the miRNAs. In sequestering the miRNAs, the ceRNAs allow the miRNA target RNAs to be expressed. According to this definition of ceRNA, the RNA products of expressed pseudogenes containing miRNA binding sites will qualify as ceRNAs. Likewise, lncRNA can act as ceRNA as well. For example, *linc-MD1* is a validated cytoplasmic lncRNA expressed during myoblast differentiation; it acts as a ceRNA for miR-133 and miR-135 targets. Phosphatase and tensin homolog (*PTEN*) is a tumor suppressor gene whose expression is frequently altered in many human cancers. The regulation of *PTEN* expression by a whole plethora of miRNAs is further modulated by ceRNAs, such as VAPA and CNOT6L.[25]

The **circular RNAs** (**circRNAs**) with a functional role are the latest addition to the RNA universe. The existence of RNAs in circular form at a low level had been reported earlier; these were treated as unique, sporadic observations. The extensiveness of circRNA expression was reported in 2012.[26] The authors concluded that a non-canonical mode of RNA splicing, resulting in a circular RNA isoform, is a general

feature of the gene-expression program in human cells, and that the expression of circRNAs is more prevalent and widespread than once thought. However, the regulatory role of circular RNAs was highlighted by two recent publications.[27,28] Both these publications described highly stable circular RNAs in human and mouse brain (termed CDR1as, for antisense (as) to the cerebellar-degeneration-related protein 1 transcript CDR1, by Memczak et al., and ciRS-7 for circular RNA sponge for miR-7, by Hansen et al.). These circRNAs bind many copies of miR-7 and terminate miR-7-mediated suppression of target mRNAs. These circular RNAs contain approximately 70 conserved binding sequences for miR-7. Overexpression of this circRNA reversed the miR-7-mediated suppression of the target mRNAs; hence, expressing this circRNA or deleting the miR-7 had the same phenotypic outcome. Hansen et al. also reported that the testis-specific circRNA *Sry* (sex-determining region Y) serves as a miR-138 sponge.

The existence of the different forms of noncoding regulatory RNAs makes sense from the standpoint of building robustness in the regulatory network. However, it is tempting to speculate that the coexistence of various forms of noncoding RNAs may also determine the degree of titration needed to reach the threshold of effects in a cell-specific manner.

1.9 PROTEIN STRUCTURE AND FUNCTION

Proteins (polypeptides) are translated from the mRNA, which carries the amino acid sequence information for the polypeptide. Translation proceeds from the N-terminal to C-terminal direction of the polypeptide being synthesized. Proteins are made up of structural units called amino acids. All amino acids are **α-amino acids**. They are called α-amino acids because the amino group ($-NH_2$) is attached to the α-carbon atom—that is, the carbon atom linked to the carbonyl carbon of the carboxyl group (−COOH). The basic formula of an amino acid is shown in Figure 1.6A.

1.9.1 Configuration and Chirality of Amino Acids

All amino acids except glycine (R = H) are **chiral** because the α-carbon is chiral or asymmetric. So, **except for glycine** all amino acids can have two mirror-image stereoisomers (enantiomers). According to the DL system of Fischer, all natural amino acids are in L-configuration (as opposed to monosaccharides, which exist in D-configuration) (Figure 1.6B); according to the RS system of Cahn−Ingold−Prelog, all natural amino

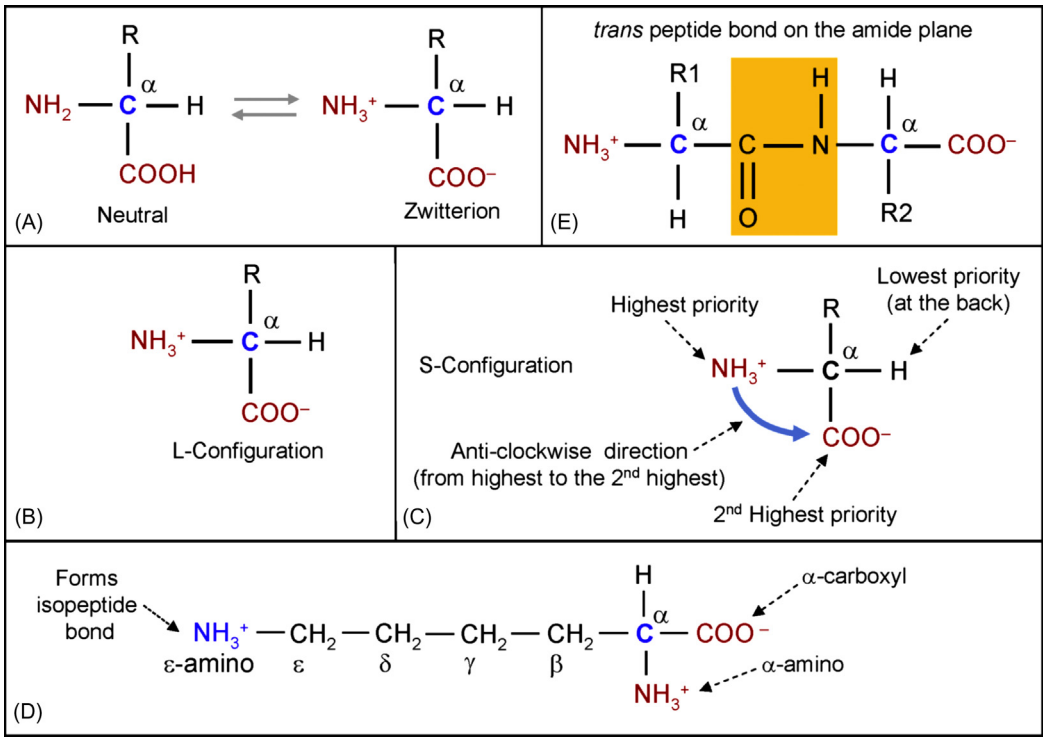

FIGURE 1.6 **Amino acid structure and peptide bond.** All amino acids except glycine (in which R = H) are chiral because the α-carbon is asymmetric. (A) Basic formula of amino acids; (B) L-configuration of amino acid per Fischer's system; (C) S-configuration of amino acid per Cahn−Ingold−Prelog rules; (D) the numbering of carbon atoms for lysine; (E) the peptide bond is a *trans* bond on the amide plane (in color).

BOX 1.8

1. The DL system of denoting enantiomers, originally introduced by Emil Fischer, is an old way of denoting the chirality of biological macromolecules. A more recent system is the RS system introduced by Robert Cahn, Christopher Ingold, and Vladimir Prelog. Naturally occurring amino acids have L-configuration according to the DL system, and S-configuration according to the RS system. In the RS system, first the priority of the groups attached to the chiral center is established. Then the order from the highest priority group to the second highest priority group, and so on, is established. If the order is clockwise, the molecule is said to have the R- (rectus) configuration; if the order is anticlockwise, the molecule is said to have S- (sinistrus) configuration. In Figure 1.6, NH_3^+ has the highest priority (because the atomic number of N is 7), followed by COO^- (because the atomic number of C is 6). If the first atom of two groups has the same atomic number, then the priority of the group is determined by the second atom and so on. Thus, COOH will have higher priority than CH_2OH.
2. The presence of two H atoms makes the α-carbon of glycine achiral (not chiral) or symmetric. As a result, glycine does not have any enantiomer (D/R or L/S isomer) and has no optical activity (dextro or levo).

acids are in the S-configuration (Figure 1.6C). So, the S-form is analogous to the L-form (see Box 1.8). Located on the alpha carbon is the "R" group, called the **side chain**. The nature of this side chain determines the identity of a particular amino acid. Glycine is the simplest amino acid because R = H. Amino acid side chains can be polar or nonpolar. Polar side chains may be charged or neutral. For example, two negatively charged amino acids are aspartic acid and glutamic acid. Two positively charged (i.e. protonated) amino acids are lysine and arginine. Figure 1.6D shows the numbering of carbon atoms of lysine. A small fraction of histidine is also positively charged at physiological pH. Proline is the only amino acid that has an imino group rather than an amino group. Although there are many more amino acids known so far, only 20 of them are standard amino acids used by all organisms during translation to synthesize proteins because they are encoded by the genetic code.

1.9.2 Ionic Character of Amino Acids

In solution at physiological pH (7.4), amino acids exist as dipole ions or **zwitterions**, where the amino group (NH_2) exists as an ammonium ion (NH_3^+) and the carboxyl group (COOH) exists as a carboxylate ion (COO^-) (Figure 1.6A). An amino acid can therefore act as a base as well as an acid, and hence is an ampholyte (having amphoteric properties). In a zwitterion, the + and − charges cancel each other to give the molecule a net charge of zero. However, at pH that is significantly higher or lower than physiological pH, amino acids undergo ionization. At acidic pH that is significantly lower than 7.4, the amino group has a positive charge while the carboxyl is neutral. At alkaline pH that is significantly higher than 7.4, the amino group is neutral while the carboxyl has a negative charge.

Amino acids of proteins in solution accept or lose protons depending on the nature of the side chains. The pK_a values of amino acids (i.e. the tendency of amino acids to lose protons) play an important role in determining the pH-dependent properties of a protein in solution. Internal ionizable groups in proteins are essential for catalysis. During a cycle of function, these internal ionizable groups can experience different microenvironments, and their pK_a values and charged states adjust accordingly.[29]

1.9.3 Relationship between Protein Function and the Location of Amino Acids in the Polypeptide Chain

The location of amino acids in the folded conformation of a protein is relevant for the protein's function and its interaction with the environment. For example, proteins located in a hydrophobic environment, such as membrane, have nonpolar (hydrophobic) side chains on the surface interacting with the membrane lipids. In contrast, proteins located in an aqueous environment, such as cytosol, have polar side chains (hydrophilic) on the surface interacting with the aqueous environment.

Arginine and lysine carry positive charges, and are often located on the interacting surface of proteins that interact with negatively charged molecules. Predictably, arginine and lysine are found on the surface of DNA-binding proteins that interact with the negatively charged phosphate group of DNA.

Similarly, aspartic acid and glutamic acid carry negative charges, and are often located on the interacting surface of proteins that interact with positively charged molecules. Aspartic acid and glutamic acid in calmodulin bind Ca^{++} ions, which carry a complementary positive charge. Many proteins in halophilic archaebacteria, which live in an extremely salty environment, have high localized concentrations (high charge density) of acidic amino acids on the surface. Such high charge density of acidic amino acids very effectively sequesters sodium ions, thus preventing denaturation and precipitation of cellular proteins. In fact, these proteins are denatured if placed in low salt concentration because the removal of sodium ions leaves many closely placed negative charges exposed, which strongly repel each other.

Serine, threonine, and tyrosine have hydroxyl groups (−OH) in their side chains. These OH groups can serve as phosphate attachment sites during phosphorylation. Many receptors that are involved in signal transduction are phosphorylated for activation, and consequently have these amino acid residues in their active sites. Phosphorylation causes conformational change in these receptors.

The sulfhydryl (-SH) group in cysteine is ideal for binding metals through metal—thiolate bonds. Naturally, cysteines are prevalent in many storage proteins that bind heavy metals. For example, in metallothionein, the intracellular metal-binding protein, one third of the amino acid residues are cysteines. The -SH group is also ideal for forming strong covalent disulfide linkages that stabilize the conformation of proteins. Expectedly, cysteines are found in many enzymes that function in harsh conditions of salt and pH, such as digestive enzymes like pepsin and chymotrypsin. The structure of many small proteins, such as insulin and ribonuclease, is stabilized by cysteine disulfide linkages. Cysteine disulfide linkages also confer rigidity to protein tertiary structure and are found in proteins like keratin in hair.

Proline occurs near the bend of polypeptide chains, and its ring forms a useful kink in the protein chain. Therefore, proline helps redirect the protein chain back inwards or around a tight corner.

Glycine and alanine, being very small, are flexible and can easily fit into tight spots. For example, glycine is the most abundant amino acid in the tight triple helix of collagen (about one-third of all amino acids). Alanine, being small and chemically inconspicuous, can be accommodated on the inside as well as outside of proteins. Alanine residues are very common in proteins. Attempts to confirm the functional role of specific amino acid residues in proteins involve mutagenesis experiments, and oftentimes the target amino acid is replaced by alanine.

1.9.4 Linkage between Amino Acids—The Peptide Bond

Amino acids are linked together by peptide bonds (**alpha peptide bonds**), which are simply **amide linkages** between the NH_2 and COOH groups of neighboring amino acids. The peptide bond has unique characteristics, which contribute to the overall structure of proteins. The peptide bond has a partial double-bond character. Thus, it is rigid and planar and not free to rotate. The plane on which it lies is called the **amide plane**. Peptide bonds are generally ***trans*** bonds—that is, the carbonyl oxygen and amide hydrogen are in *trans* position (Figure 1.6E). The Cα−C bonds are not rigid and they can freely rotate, being only limited by the size and character of the R groups. In lysine, the ε-amino group (Figure 1.6D) also participates in the formation of a peptide bond, which is called an **isopeptide bond** because it does not involve the usual α-amino group.

1.9.5 Four Levels of Protein Structure

Proteins have four levels of structure: primary, secondary, tertiary, and quaternary. **Primary structure** refers to the amino acid sequence of a protein. **Secondary structure** refers to the conformation of the polypeptide backbone. Examples of secondary structures are helices (**α-helix**), pleated sheets (**β-pleated sheet**), and bends or turns (**β-bend**). **Tertiary structure** of a protein refers to its three-dimensional structure—that is, further folding of the secondary structure in the three-dimensional space. **Quaternary structure** refers to a structure achieved by proteins composed of more than one polypeptide chain. Each polypeptide chain, called a subunit, has its own primary, secondary, and tertiary structure. In quaternary structure, protein chains (subunits) can associate with one another to form dimers, trimers, and other higher orders of oligomers. Recent studies have shown that despite having definitive structure, many proteins have specific regions that are intrinsically disordered (see Box 1.9).

1.9.6 Acidic and Basic Proteins

At physiological pH (7.4), acidic proteins tend to be negatively charged and have a higher proportion of acidic amino acids (e.g. aspartic acid, glutamic acid), whereas basic proteins tend to be positively charged and have a higher proportion of basic amino acids (e.g. arginine, lysine).

Hydrophilic and charged amino acids are frequently associated with antigenic determinants (**epitopes**),

BOX 1.9

INTRINSICALLY DISORDERED PROTEINS: THE "UNSTRUCTURAL" ASPECT OF STRUCTURAL BIOLOGY[30]

It has long been known that structural flexibility exists in proteins and aids in ligand binding. Nevertheless, the "structure–function paradigm"—that is, that proteins possess definitive three-dimensional structures in order to perform their function—has been the standard paradigm in protein biochemistry. Experimental evidence accumulating since the turn of the millennium has brought to light a unique aspect of protein structure that challenges this traditional structure–function paradigm once thought to be a universal theme applicable to all proteins. These findings demonstrate that under native functional conditions, many proteins or specific regions of some proteins are intrinsically disordered, existing as molten globules, collapsed or extended random coils, transiently structured forms, etc. These proteins are called **intrinsically disordered proteins** (**IDP**s). IDPs lack a unique three dimensional structure, either entirely or in part, when alone in solution. About 10–35% of prokaryotic and about 15–45% of eukaryotic proteins are estimated to contain disordered regions that are at least 30 amino acid residues in length. A significant number of IDPs are involved in regulatory and signaling functions; hence, IDPs are more prevalent in eukaryotes than in prokaryotes. IDPs and IDP databases are discussed in section 8.11 (Chapter 8).

such as arginine, lysine, aspartic acid, glutamic acid, asparagine, glutamine, serine, and threonine.

1.9.7 Nonstandard Amino Acids in Polypeptide Chains

As indicated earlier, selenocysteine and pyrrolysine are the two nonstandard amino acids that are incorporated directly into the polypeptide chain during translation. Selenocysteine has been found in lower as well as higher organisms (including mammals), while pyrrolysine has so far been found in certain archaebacteria. However, their occurrence in proteins is not nearly as universal as the 20 standard amino acids.

1.10 GENOME STRUCTURE AND ORGANIZATION

The genomic DNA in the nucleus exists in combination with **histone** proteins; the DNA–protein complex is known as **chromatin**. The unit of chromatin is the **nucleosome**; thus, chromatin can be envisioned as a repeat of regularly spaced nucleosomes. A nucleosome core particle is composed of a **histone** octamer and the DNA that wraps around the octamer (Figure 1.7). Histones are globular basic proteins with a flexible N-terminal end (the so-called "tail") that is subject to various covalent modifications (epigenetic modifications). The histone octamer is composed of two molecules each of histones H2A, H2B, H3, and H4. DNA wraps around the octamer in a left-handed supercoil of about 1.75 turns that each contain approximately 150 bp. Histone H1 is the **linker histone** that, along with **linker DNA**, physically connects the adjacent nucleosome core particles. Each nucleosome has a diameter of 10 nm, and the nucleosomes are compacted into a solenoid fiber structure of 30 nm (see Box 1.10). The 30-nm solenoid fibers undergo further progressive compaction into 300-nm filament, and ultimately into a 700-nm chromosome. During cell division, when the chromosomes duplicate, a 1400-nm metaphase chromosome is produced, containing two chromatids, each chromatid being 700 nm (Figure 1.7).

The major non-histone proteins associated with chromatin are the **high mobility group (HMG**) proteins. Whereas histones increase the compactness of the chromatin, HMG proteins decrease its compactness. By decreasing the compactness of the chromatin, HMG proteins facilitate the accessibility of various regulatory factors to DNA. HMG proteins can also bind to DNA and cause significant bending of the DNA. DNA bending is important for the interaction between transcription factors and **coregulators (coactivators/corepressors**[d]) in regulating transcription.

Various protein–DNA interactions can make the chromatin undergo changes in its conformation in response to various cellular metabolic demands. Altered chromatin conformation, in turn, can limit or enhance the accessibility and binding of the transcription machinery, thereby regulating transcription. Some of these regulatory effects could be mediated epigenetically.

[d]Coactivators and corepressors are proteins that do not bind DNA themselves, but interact with DNA-binding proteins, to either upregulate or downregulate transcription.

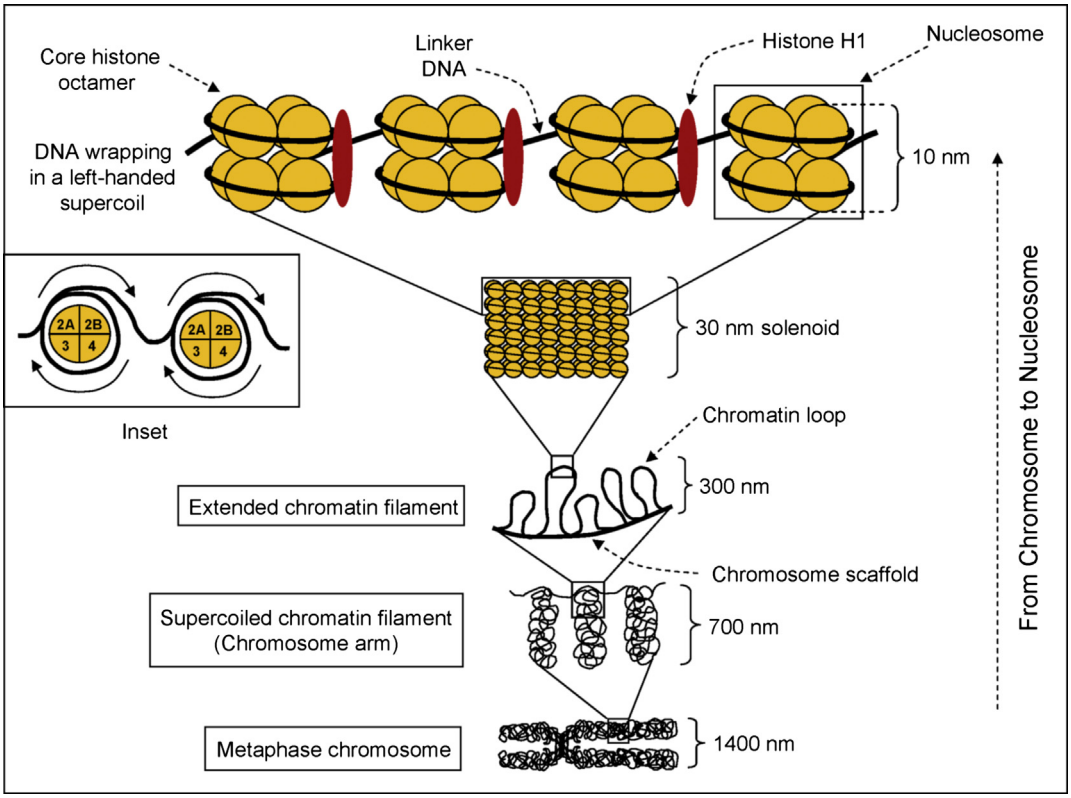

FIGURE 1.7 **The hierarchy of organization from chromosome to nucleosome.** Inset shows the relative position of histone monomers with respect to one another and the direction of wrapping of DNA around nucleosomes. *(Figure reproduced from Choudhuri et al. (2010) Toxicol. Appl. Pharmacol. 245: 378–393, with some modifications.)*

BOX 1.10

CHROMATIN FIBERS: 30 NM OR 10 NM?

Figure 1.7 shows the prevailing model of genome organization, which is the subject of textbooks. This model has been in existence since the mid-1970s, and it describes chromatin as a 30-nm fiber, which is formed by the coiling of the basic 10-nm fiber. Recent experimental evidence has challenged this traditional concept of chromatin organization.[31] By combining electron spectroscopic imaging with tomography, the authors generated a three-dimensional image that revealed that both open and closed chromatin domains in mouse somatic cells comprise 10-nm fibers. This indicates that the 30-nm chromatin model does not reflect the true regulatory structure in vivo. So, why was chromatin fiber reported to be 30 nm? This puzzle remains to be solved to the satisfaction of chromatin biologists. It has been suggested that it could be a combination of methodological artifact associated with chromatin isolation, as well as the inability to detect and distinguish the existence of the 10-nm fibers in the background of 30-nm fibers. Additional studies are expected to resolve this issue in the near future.

1.10.1 The Structure of a Representative Genome—The Human Genome

The human genome is discussed here as the representative genome.[32–34] The human genome consists of 3.2 billion (3.2×10^9) base pairs (=3.2 Gbp), distributed in 23 pairs of chromosomes (22 pairs of autosomes + XX or XY sex chromosomes). There are ~21,000 protein-coding genes, and the protein-coding fraction of the DNA constitutes ~1.5–2% of the entire genomic DNA. About two-thirds of the protein-coding genes have 1:1 **orthologs**[e] across placental mammals. Regulatory

[e]Genes in different species but related by speciation events are called **orthologous genes** or **orthologs**. Depending on the number of genes found in each species, the relationship of orthologs could be 1:1, 1:many, and many:many.

sequences constitute ~3–3.5% of the genome. The genome also codes for a significant number of noncoding regulatory RNAs. Initial studies suggested that more than 10% of the genome is represented in mature transcripts, and ~20% of the genome may be functionally important. These estimates have been revised and significantly expanded based on the findings of the Encyclopedia of the DNA Elements (ENCODE) project, discussed later. *The genomes of two humans are about 99.9% identical*.

Repeat sequences account for ~50% of the human genome; hence repeat sequences constitute a significant source of genetic diversity. Repeat sequences are of various types: **simple repeats** (e.g. $(A)_n$, $(CA)_n$, $(CGG)_n$), **tandem repeat blocks** (e.g. centromeric repeats, telomeric repeats, ribosomal gene clusters), **segmental duplications** (e.g. blocks of 1–200 kb or longer repeats copied from one region of the genome and integrated into another region of the genome), **interspersed repeats** (transposable-element-derived), and **processed pseudogenes**. In addition to the repeat content, further functional genetic diversity is imparted by **single nucleotide polymorphism** (**SNP**) and **copy number variation** (**CNV**), also called **copy number polymorphism** (**CNP**). According to older definition, a point mutation has to occur in at least 1% of the population in order to qualify as an SNP, but this is no longer strictly followed; all point mutations are called SNPs. In the human genome, >65% of all SNPs are C→T transition mutations.

Recent evidence suggests that the human genome is extensively transcribed. However, the fraction of the genome that is transcribed into functional noncoding transcripts is yet to be estimated precisely. The findings from the **Encyclopedia of the DNA Elements** (**ENCODE**) project suggest that the noncoding yet functional fraction of the genome may vary significantly from chromosome to chromosome. There is also evidence for both sense and antisense transcription in the human genome. There is extensive alternative splicing of transcripts so that there are well above 100,000 proteins encoded by the human genome.

The G + C-rich regions of the genome are gene-dense, and the genes in these regions are smaller and more compact due to smaller intron size. Conversely, A + T-rich regions are gene-poor and the genes in these regions are longer because of longer intron size. Average G + C content of the entire human genome is 41%, but local G + C contents may vary significantly. An important component of the G + C-rich genomic regions is the CpG sequence, which may or may not occur in clusters. CpG clusters are called **CpG islands**. The human genome contains about 0.8% CpG islands. However, based on the G + C content (~41%), the CpG island frequency should be ~4%. The difference is due to the fact that the cytosine of the CpG island is methylated, and over evolutionary time the methyl cytosine (^{me}C) tends to spontaneously deaminate to thymine, hence converting CpG to TpG. The $^{me}C \rightarrow T$ mutation creates a T–G mismatch in the DNA double strand and is normally repaired; however, it sometimes escapes the repair machinery (e.g. if it happens before replication and strand separation). The CpG islands are associated with the 5′-ends of many genes. Identification of CpG islands thus helps define the 5′-ends of genes. Methylation of the C of CpG is associated with transcriptional silencing, and the absence of methylation is associated with active transcription. Thus, unmethylated CpG islands are associated with the promoters of transcriptionally active genes, such as housekeeping genes, and genes showing tissue-specific expression.

The birth of new genes and the death of existing genes in the genome are important events that contribute to genome evolution. New genes can be born or acquired by a genome. New genes can be born through one of multiple genomic events, such as **gene duplication**, **de novo gene origination**, and **transposable element (TE) domestication**. Duplicated genes can diverge and acquire new function. These genes are called **paralogous genes** or **paralogs**[f]. New genes can be born de novo by functionalization of a previously noncoding region of the DNA. Sometimes genomes can recruit TEs and use the TE-encoded protein as the cellular protein. New genes can also be acquired through **lateral gene transfer**. Genome evolution is discussed in more detail in Chapter 2.

Gene death occurs when genes acquire inactivating mutations and lose function. Pseudogenization is a common mechanism of gene death. Pseudogenes may be **non-processed pseudogenes** or **processed pseudogenes**. Non-processed pseudogenes are an inactivated form of a gene that has acquired inactivating mutations; hence they may have intact exon–intron organization but the ORF is disrupted. In contrast, processed pseudogenes result from the reverse transcription of mRNA into complementary DNA (cDNA), followed by the integration of the cDNA into the genome. Thus, processed pseudogenes may have a poly(A) tail but they lack a promoter and other 5′-regulatory elements. (see Box 1.11)

[f]**Paralogous genes** or **paralogs** are produced through gene duplication within a genome. Paralogs may evolve new functions or may become pseudogenes.

BOX 1.11

More than a decade after genome sequencing, we are still far from understanding many aspects of structural and functional genomics, such as the exact number of protein-coding and non-protein-coding genes and their genomic locations; the genome-wide distribution of functional regulatory elements; the regulation and coordination of gene expression at different levels and regulation of the regulators; chromatin dynamics; epigenetic editing of the language of DNA; gene and protein networks; protein−protein interactions; regulation of interaction specificity in biological systems and the specificity determinants, such as protein interaction specificity and signaling specificity; the correlation between genetic diversity and disease susceptibility; the molecular determinants of humanness, that is, what it means to be a human at the molecular level; and many such similar questions.

1.10.2 Functional Sequence Elements in the Genome

Functional sequence elements of the genome regulate genome expression. These are promoters, enhancers, silencers, locus control regions (LCRs), and insulators. Elements that aid in the termination of transcription (terminators) are not discussed here.

1.10.2.1 Promoters

The 5′-flanking region of the gene is the region upstream of the transcription start site (+1). It contains the promoter and other *cis*-acting transcription regulatory sequence elements. A **promoter** is a *cis*-acting transcription regulatory element that initiates the transcription of a gene. The various regions of the promoter are termed the **core** (or **basal**) promoter, **proximal** promoter, and **distal** promoter, based on their distance from the transcription start site. Typically, the core promoter is about 35 bp long, and can extend between the −35- and +35-nt position (with respect to the +1 site). The core promoter may contain two or more of the following sequence motifs: **TATA box**, **initiator** (**Inr**) **element**, and **downstream promoter element** (**DPE**). Upstream of core promoter is the proximal promoter, which is about 250-bp long and can extend between the −250 and +250-nt position. However, in the literature, sequences far upstream of −250 have also been referred to as proximal promoter sequences. Sequences that are further upstream of the proximal promoter elements are called the distal promoter. In general, the transcription start site is determined by the TATA box and the initiator element, or in the case of TATA-less promoters, by the initiator element and the downstream promoter element, all located within the core promoter.[18]

1.10.2.2 Enhancers

Enhancers bind specific transcriptional activators and enhance the rate of transcription. Enhancers can be located close to the transcription start site, upstream or downstream from the transcription start site, and even within introns. An enhancer can regulate more than one gene in a position- and orientation-independent manner. The mechanism of enhancer action is thought to involve looping of the DNA, thereby bringing the enhancer-bound transcriptional activators close to the promoter-bound transcription factors. In doing so, enhancers increase the concentration of activators near the promoter, which directly or indirectly interact with the promoter to initiate transcription. The interaction of enhancer-bound transcriptional activators and promoter-bound transcription factors is mediated by coactivators. **Coactivators** are proteins that do not bind DNA themselves but interact with DNA-bound transcriptional activator proteins, thereby facilitating protein−protein interaction. Some examples of coactivator proteins are CBP/p300, p160, p300/CBP-interacting protein (p/CIP), p300/CBP-associated factor (p/CAF), yeast transcriptional adaptor GCN5, steroid receptor coactivator-1 (SRC-1), and there are many others. The opposite of enhancers are **silencers**, which bind transcriptional suppressor proteins and suppress transcription, thereby acting as negative regulatory elements. Like enhancers, silencers can also function in an orientation-, position-, and distance-independent manner, and they can also be located within introns.

1.10.2.3 Locus Control Regions

A locus control region (LCR) enhances the transcription of a cluster of linked genes by inducing a more open conformation of the chromatin flanking the locus. The LCR of the human β-globin locus has been well studied. The transcription-enhancing activity of LCRs is mediated by the binding of specific transcriptional activator proteins. Because LCRs can induce conformational change of chromatin, they play important roles in regulating the transcriptional activity of the euchromatic regions of chromosomes.

1.10.2.4 Insulators

Insulators are gene-boundary elements; these are DNA sequence elements that, when bound to insulator-binding proteins, shield a promoter from the effects of nearby regulatory elements. There are two types of insulator functions: an **enhancer-blocking function** and a **heterochromatin barrier function**. When an insulator is located in between a promoter and an enhancer, the enhancer-blocking function of the insulator shields the promoter from the transcription-enhancing influence of the enhancer. The heterochromatin barrier function of an insulator prevents a transcriptionally active euchromatic region from turning into transcriptionally inactive heterochromatin by the inactivating effect of the invading adjacent heterochromatin[g]. An example of an enhancer-blocking insulator is the **gypsy** insulator in *Drosophila*. The chicken **β-globin insulator** (**cHS4**), which is highly rich in G + C and the most extensively studied vertebrate insulator, has both enhancer-blocking and heterochromatic barrier functions. The mechanism of the enhancer-blocking function may involve DNA looping, but it is yet to be established. However, the mechanism of heterochromatic barrier function understandably involves the maintenance of active chromatin configuration through histone modifications at the boundary. Various proteins that bind to these insulator sequences have been identified.[35]

1.10.3 Epigenetic Modifications of the Genome Can Edit the Language Written in the DNA Sequence and Add an Extra Layer of Complexity in Genome Expression

Epigenetics is the study of mitotically or meiotically heritable changes in gene function that cannot be explained by changes in the DNA sequence.[36] Epigenetic inheritance involves the transmission of epigenetic marks not encoded in the DNA sequence, from parent cell to daughter cells and from generation to generation. Epigenetic regulation of genome expression is mediated by three main mechanisms: (1) **DNA methylation**, (2) **histone modification and chromatin conformation change**, and (3) **regulation of gene expression by ncRNAs**. DNA methylation involves the covalent addition of a methyl group to the carbon-5 position of cytosine to form 5-methylcytosine (5-mC) in CpG dinucleotides. Methylation is catalyzed by three major DNA methyltransferases (DNMTs), and the methyl group donor is S-adenosylmethionine (SAM). The de novo methylation establishes the parent-specific methylation pattern, and maintenance methylation replicates the methylation pattern of the parent strand to the daughter strand during DNA replication. This is accomplished by first recognizing the hemi-methylated CpG sites at the replication foci, followed by the addition of methyl groups to cytosines on the nascent DNA strand to re-establish the parent-specific methylation pattern. The de novo methyltransferases are DNMT3A and DNMT3B, whereas the maintenance methyltransferase is DNMT1.

Methylation of the C of CpG is associated with transcriptional silencing, and the absence of methylation is associated with active transcription. Thus, unmethylated CpG islands are associated with the promoters of transcriptionally active genes, such as housekeeping genes and genes showing tissue-specific expression. Transcriptional silencing by DNA methylation is mediated by a condensed state of chromatin. Conversely, transcriptionally active genes maintain an open state of chromatin.

Covalent histone modification—such as acetylation, methylation, phosphorylation, ubiquitination, or sumoylation of specific amino acid residues, such as lys (K), arg (R), ser (S) and others, but mainly lys residues of different histone subunits—can either upregulate or downregulate gene expression. All known histone acetylation and phosphorylation modifications are transcription-activating, whereas all known sumoylations are transcription-silencing. Histone methylation and ubiquitination can be transcription-activating or silencing, depending on the specific residue modified. Table 1.2 shows some transcriptional-activating and repressing histone modifications. Epigenetic orchestration of genome expression is a tightly regulated process and it involves the cross-talk between DNA methylation and histone modifications.[37]

Regulation by small ncRNAs (e.g. miRNAs, siRNAs) is another means of epigenetic regulation of gene and genome expression. Small ncRNA-mediated silencing of gene expression, known as **RNA interference** (**RNAi**), is achieved either by translational repression (by miRNA) or by mRNA degradation (by siRNA).[22]

Some of the relatively well studied examples of epigenetic phenomena regulating gene and genome expression are **transvection** (observed in dipteran insects), **genomic imprinting**, **X-chromosome inactivation**, **paramutation**, and **heterochromatin spread and position effect variegation**.[38] Although epigenetic mechanisms can edit the language of DNA written in its

[g]Sometimes, indiscriminate propagation of heterochromatin into adjacent euchromatin results in silencing of genes located in close proximity to the propagating heterochromatin. The silencing is often not complete; the genes are silenced in some cells, but in other cells they are expressed, resulting in a so-called variegated (patchy) expression pattern. Because this expression pattern is brought about by the proximity of the genes to the heterochromatin, the phenomenon is called **position-effect variegation (PEV)**.

TABLE 1.2 Some Transcription-Activating and Repressing Histone Modifications

Some Transcription-Activating Modifications
Acetylation
Histone H2A: K5, K9, K13; **Histone H2B:** K5, K12, K15, K20; **Histone H3:** K9, K14, K18, K23, K56; **Histone H4:** K5, K8, K13, K16
Phosphorylation
Histone H3: T3, S10, S28, Y41; **Histone H2AX:** S139 (for DNA repair)
Methylation (me1/me2/me3)
Histone H3: K4, K9 (me1), K36, K79, R17, R23; **Histone H4:** R3
Ubiquitination
Histone H2B: K120, K123 (yeast)
Some Transcription-Silencing Modifications
Methylation (me1/me2/me3)
Histone H3: K9 (me2, me3), K27; **Histone H4:** K20
Ubiquitination
Histone H2A: K119
Sumoylation
Histone H2A: K126 (yeast); **Histone H2B:** K6, K7 (yeast); **Histone H4:** K5, K8, K12, K16, K20

base sequence, thereby altering genome expression, epigenetic modulation of gene and genome expression needs further characterization. For example, much needs to be understood in terms of the correlative versus causal effects between exposure to various environmental factors and epigenetic changes. Additionally, we are not yet able to distinguish between adaptive and adverse epigenetic changes. Normal epigenetic changes associated with age and different life stages need to be thoroughly characterized as well. Some preliminary data are available but more work is underway.

1.10.3.1 Histone Code

Strahl and Allis[39] coined the term **histone code** to describe the concept that specific histone modifications could act sequentially or in combination to form a recognizable "code" that could regulate transcription as well as the state of chromatin condensation. Turner[40] used the term **epigenetic code**, which was conceptually same as the histone code. For example, phosphorylation of histone H3 serine 10 (H3S10) stimulates acetylation of histone H3 lysine 14 (H3K14), which is a transcription-activating modification; monoubiquitination of histone H2B lysine 120 (H2BK120) stimulates methylation of histone H3 lysine 4 (H3K4), which is also a transcription-activating modification.[41] See Box 1.12 regarding symmetrical and asymmetrical histone code.

BOX 1.12

ASYMMETRICAL MODIFICATION OF HISTONE AND ASYMMETRICAL HISTONE CODE

The traditional view assumes that histone code is symmetrical; that is, both molecules of the same histone in a nucleosome are modified in the same way. However, recent experimental evidence challenges this long-held view.[42] Using preparations of chromosomal mononucleosomes from embryonic stem cells, mouse embryonic fibroblasts, and cultured HeLa cells, the authors showed the existence of di- and trimethylation of lysine 27 of histone H3 (H3K27me2/3) both symmetrically and asymmetrically in native chromatin in approximately equal proportions. When the H3K27me2/3 mark occurred asymmetrically there was a different methylation mark on the sister histone, either H3K4me3 or H3K36me2/3. In other words, in a nucleosome, one of the two H3 molecules contains one mark, while the other H3 contains a different mark. Whereas H3K4me3 or H3K36me2/3 are transcription-activating modifications, H3K27me2/3 is transcription-repressing modification. The coexistence of such antagonizing histone modification marks might facilitate rapid and efficient regulation of transcription because the removal of one of these marks may be sufficient to rapidly induce transcriptional activation or repression. The existence of asymmetric histone modifications also shows that histone code could be symmetric or asymmetric. The possibility of existence of asymmetric histone modification marks throughout the genome significantly expands the scope of epigenetic regulation, particularly when the combinatorial aspect of such modifications and their effect on transcription are taken into account.

1.10.3.2 The Dynamics of Epigenetic Changes

Epigenetic modifications, particularly DNA methylation, have been traditionally regarded as static modifications. Progress in epigenetics during the past few years has demonstrated that epigenetic modifications of the genome are lot more dynamic than initially thought. A recent study in mice[43] suggests that epigenetic modifications can even control circadian rhythms of gene expression, thereby regulating circadian-rhythm-driven physiological processes. The authors observed circadian oscillations of several antisense RNA, long noncoding RNA, and microRNA transcripts coupled with rhythmic histone modifications in promoters, gene bodies, or enhancers in adult mouse livers. Promoter DNA methylation levels were relatively stable. The authors identified a set of 1262 (9% of expressed) oscillating transcripts, of which 1160 were protein-coding, including genes implicated in metabolic regulation, such as *Arntl*, *Cry1*, *Per1*, *Per2*, *Per3*, *Rorc*, *Foxo3*, and many others. The five investigated histone modifications—H3K4me1, H3K4me3, H3K9ac, H3K27ac, and H3K36me3—were enriched in actively transcribed genes and correlated with transcript levels. The oscillating expression of an antisense transcript (*asPer2*) to the gene encoding the circadian oscillator component *Per2* was also identified. Robust transcript oscillations often accompanied rhythms in multiple histone modifications and recruitment of multiple chromatin-associated clock components. The findings of this study, as well as some other studies before it, demonstrate that epigenetic modifications could be very dynamic and may even control rapid and short-term regulation of gene expression.

1.10.4 Lessons Learned from the Second Phase of the ENCODE Project about the DNA Elements in the Human Genome and its Epigenetic Modifications

The Encyclopedia of DNA Elements (ENCODE) project has been a logical continuation of the big science that was launched with the human genome sequencing project. ENCODE aims to delineate all functional elements encoded in the human genome. *A functional element is defined as a discrete genome segment that either encodes a product (e.g. protein or noncoding RNA) or displays a reproducible biochemical signature (e.g. protein binding, or a specific chromatin structure).* Following the initial success of the first phase of ENCODE, initiated in 2003 to characterize 1% of the human genome, the scope of ENCODE has been broadened since 2007 to study DNA elements in the whole human genome. The work in the second phase involved integration of results from experiments involving 147 different cell types, and all ENCODE data, with other resources, such as candidate regions from genome-wide association studies (GWAS) and evolutionarily constrained regions.[44,45]

Based on the analysis, about 80% of the genome was assigned some kind of genetic function, either RNA-associated or chromatin-associated. About 95% of the genome was found to lie within 8 kb of a DNA−protein interaction, and 99% within 1.7 kb of at least one of the biochemical events measured by ENCODE. The analysis annotated 8801 small RNA and 9640 long noncoding RNA-coding loci. Greater than 62% of the genomic bases were found to be represented in > 200-nt-long RNA molecules. Most transcribed bases were found to be within annotated genes or in overlapping annotated gene boundaries; that is, in noncoding DNA. Also, 11,224 pseudogenes were annotated, of which 863 are transcribed and associated with active chromatin.

An initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features were annotated. A total of 62,403 transcription start sites were identified, of which 27,362 (44%) are within 100 bp of the 5′-end of an annotated or known transcript. The remaining regions predominantly lie across exons and 3′-UTRs, some exhibiting cell-type-restricted expression, representing possible start sites of novel cell-type-specific transcripts. The binding locations of 119 different DNA-binding proteins and a number of RNA polymerase components in 72 cell types were mapped using chromatin immunoprecipitation followed by deep sequencing (ChIP-seq); 87 (73%) were sequence-specific transcription factors. Overall, 636,336 binding regions covering 231 megabases (8.1%) of the genome were found to be enriched for regions bound by DNA-binding proteins across all cell types.

Statistical models to analyze genome-wide transcription-factor-binding data identified six different types of genomic region, based on the binding data of transcription-related factors (TRFs). These six different types of genomic region form three pairs: (1) binding-active regions (BARs) and binding-inactive regions (BIRs), (2) promoter-proximal regulatory modules (PRMs) and gene-distal regulatory modules (DRMs), and (3) high-occupancy of TRF (HOT) regions and low-occupancy of TRF (LOT) regions. Region types from different pairs may overlap. For example, DRMs are subsets of BARs, and some HOT regions overlap with PRMs and DRMs. Each of these regions, however, exhibits some unique properties. The six types of region were found to occupy from about 15.5 Mbp (equivalent to 0.50% of the human genome) to 1.39 Gbp (equivalent to 45% of the human genome) in the different cell lines. Expectedly, the distribution of BARs correlates with gene density. Also, about 70 to 80% of the HOT regions were mapped within 10 kb of annotated coding and noncoding genes.

Assay for histone modifications and variants in 46 cell types showed a great deal of variability across cell types, in accordance with changes in transcriptional activity. For example, monomethylation of lysine 4 of histone H3 (H3K4me1) was found as a mark of regulatory elements associated with enhancers and other distal elements, H3K4me2 was found as a mark of regulatory elements associated with promoters and enhancers, whereas H3K4me3 was found as a mark of regulatory elements primarily associated with promoters/transcription starts. In contrast, H3K9me3 is the repressive mark found associated with constitutive heterochromatin and repetitive elements.

In conclusion, the map created by ENCODE reveals that cell type is important. In other words, cell-type-specific regulation of genome expression in multicellular organisms might hold the key to explaining not only differential regulation of gene expression, but also the development of disease.

References

1. Graci JD, Cameron CE. *Antiviral Chem Chemotherap* 2004;**15**:1–13.
2. Segal DJ, et al. *Proc Natl Acad Sci USA* 1999;**96**:2758–63.
3. Le Doan T, et al. *Nucl Acids Res* 1987;**15**:7749–60.
4. Beal PA, Dervan PB. *Science* 1991;**251**:1360–3.
5. Pilch DS, et al. *Biochemistry* 1991;**30**:6081–8.
6. Grigoriev M, et al. *Proc Natl Acad Sci USA* 1993;**90**:3501–5.
7. Chin JY, et al. *Front Biosci* 2007;**12**:4288–97.
8. Kumar A. *Eukaryot Cell* 2009;**8**:1321–9.
9. Kay RA, et al. *Cancer Res* 2005;**65**:10742–9.
10. Hawkins JD. *Nucl Acids Res* 1988;**16**:9893–908.
11. Sterner DA, et al. *Proc Natl Acad Sci USA* 1996;**93**:15081–5.
12. Choudhuri S, et al. *Biochem Biophys Res Commun* 2000;**274**:79–86.
13. Belshaw R, Bensasson D. *Heredity* 2006;**96**:208–13.
14. Lambowitz AM, Zimmerly S. *Annu Rev Genet* 2004;**38**:1–35.
15. Chorev M, Carmel L. *Front Genet* 2012;**3**:Article 55.
16. Sauna ZE, Kimchi-Sarfaty C. *Nat Rev Genet* 2011;**12**:683–91.
17. Edwalds-Gilbert G, et al. *Nucl Acids Res* 1997;**25**:2547–61.
18. Choudhuri S. In: Choudhuri S, Carlson DB, editors. *Genomics: fundamentals and applications*. New York: Informa; 2009. p. 3–48.
19. Leontis NB, Westhof E. *RNA* 2001;**7**:499–512.
20. Abu Almakarem AS. *Nucl Acids Res* 2012;**40**:1407–23.
21. Choudhuri S. *Biochem Biophys Res Commun* 2009;**388**:177–80.
22. Choudhuri S. *J Biochem Mol Toxicol* 2010;**24**:195–216.
23. Mercer TR, Mattick JS. *Nat Struct Mol Biol* 2013;**20**:300–7.
24. Amaral PP, et al. *Nucl Acids Res* 2011;**39**:D146–51 (Database Issue)
25. Tay Y, et al. *Cell* 2011;**147**:344–57.
26. Salzman J, et al. *PLoS ONE* 2012;**7**(2):e30733.
27. Memczak S, et al. *Nature* 2013;**495**:333–8.
28. Hansen TB, et al. *Nature* 2013;**495**:384–8.
29. Isom DG, et al. *Proc Natl Acad Sci USA* 2011;**108**:5260–5.
30. Tompa P. *Trends Biochem Sci* 2012;**37**:509–16.
31. Fussner, et al. *EMBO Rep* 2012;**13**:992–6.
32. Phesant M, Mattick JS. *Genome Res* 2007;**17**:1245–53.
33. Choudhuri S. In: Choudhuri S, Carlson DB, editors. *Genomics: fundamentals and applications*. New York: Informa; 2009. p. 49–99.
34. Lander ES. *Nature* 2011;**470**:187–97.
35. Valenzuela L, Kamakaka RT. Chromatin insulators. *Annu Rev Genet* 2006;**40**:107–38.
36. Riggs AD, et al. Introduction. In: Russo VEA, et al., editors. *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor, NY: CSHL Press; 1996. p. 1–4.
37. Choudhuri S, et al. *Toxicol Appl Pharmacol* 2010;**245**:378–93.
38. Choudhuri S. In: Choudhuri S, Carlson DB, editors. *Genomics: fundamentals and applications*. New York: Informa; 2009. p. 101–28.
39. Strahl BD, Allis CD. *Nature* 2000;**403**:41–5.
40. Turner B. *Bioessays* 2000;**22**:836–45.
41. Choudhuri S. *Toxicol Mech Methods* 2011;**21**:252–74.
42. Voigt P, et al. *Cell* 2012;**151**:181–93.
43. Vollmers C, et al. *Cell Metab* 2012;**16**:833–45.
44. The ENCODE Project Consortium. *Nature* 2012;**489**:57–74.
45. Yip KY, et al. *Genome Biol* 2012;**13**:R48.