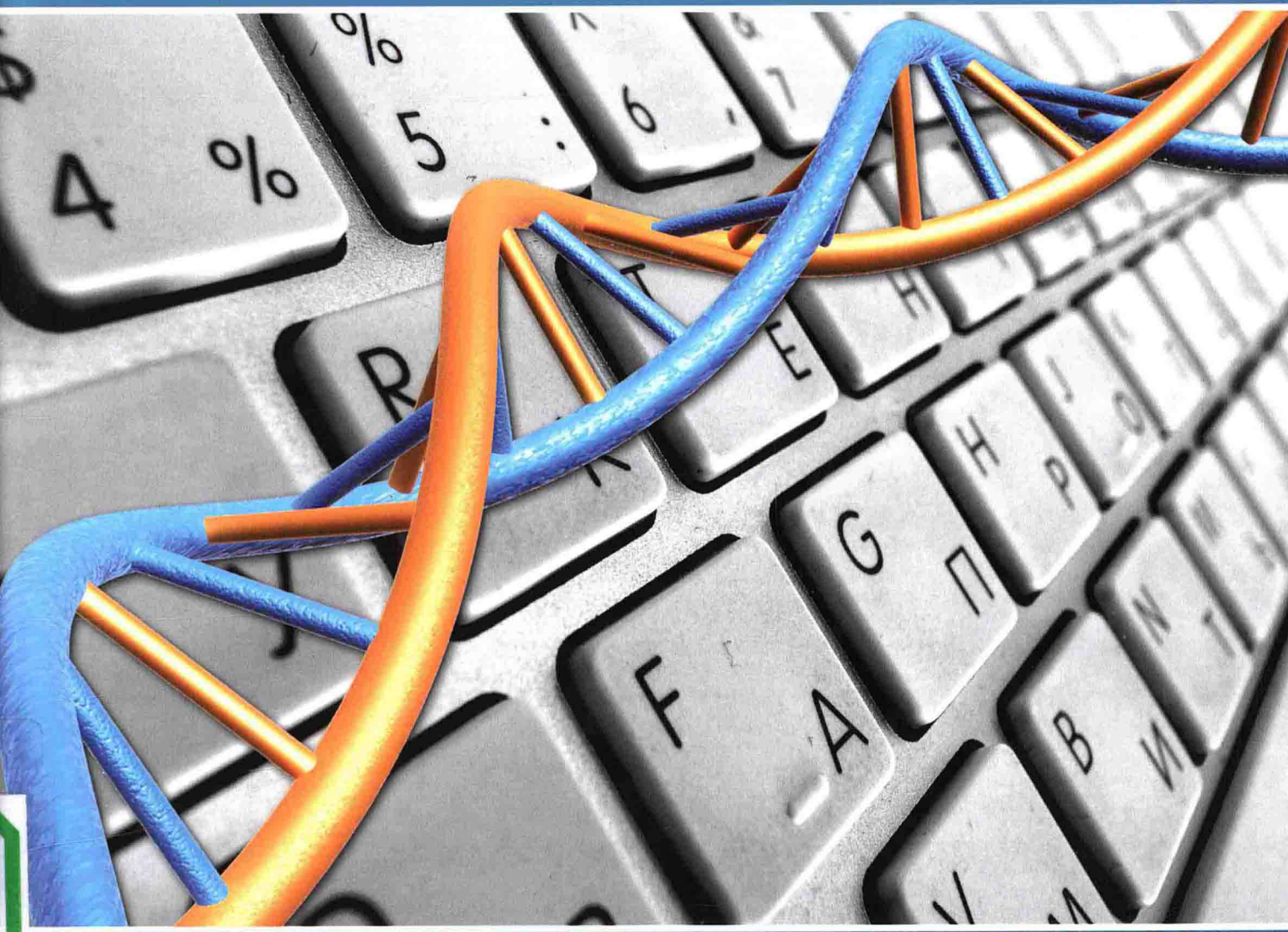
Bioinformatics for Beginners

Genes, Genomes, Molecular Evolution, Databases and Analytical Tools



Supratim Choudhuri

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

BIOINFORMATICS FOR BEGINNERS

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.

Contents

7. Additional Bioinformatic Analyses Involving Nucleic-Acid Sequences

8. Additional Bioinformatic Analyses Involving Protein Sequences

9. Phylogenetic Analysis

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.

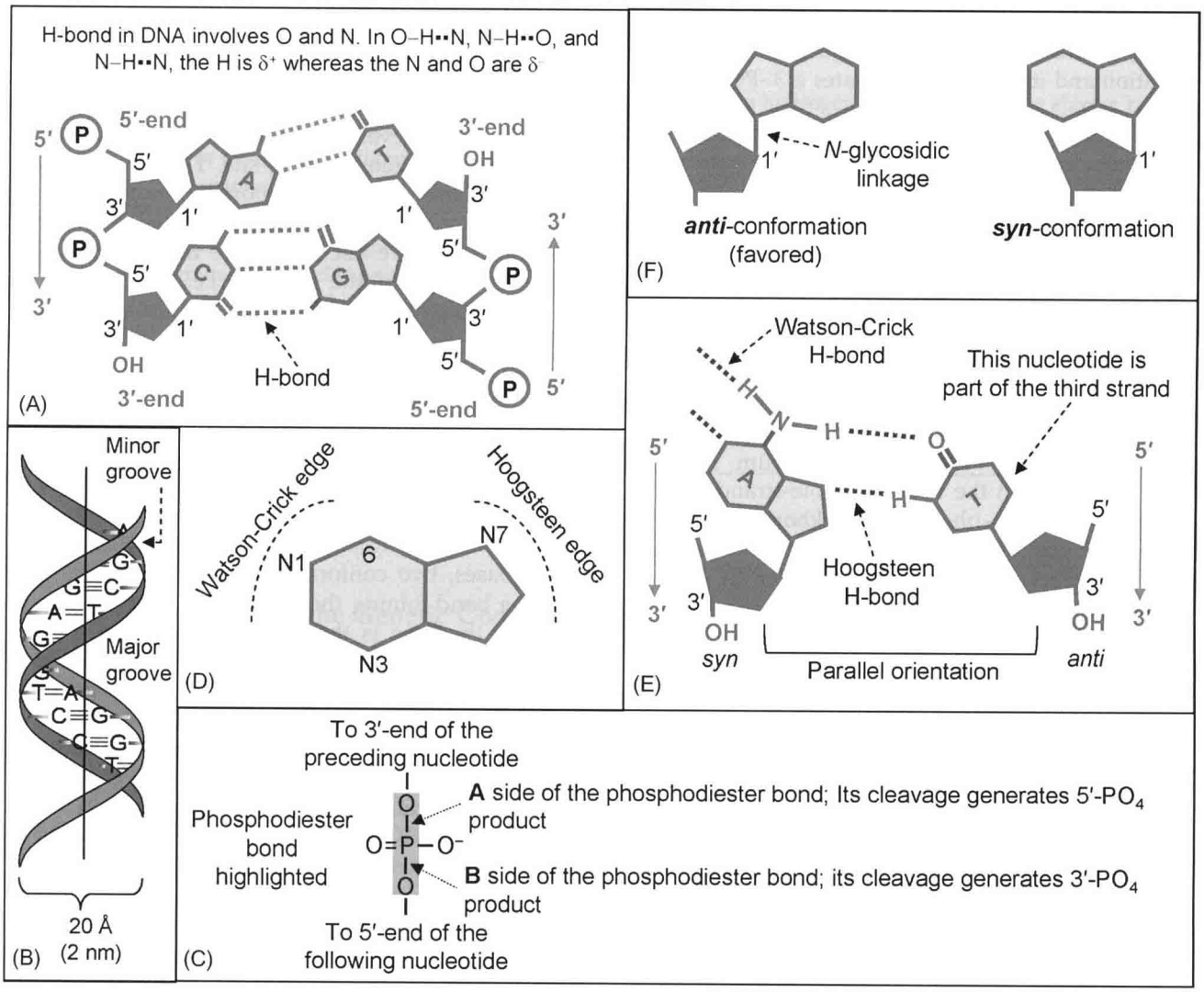


FIGURE 1.1 **DNA structure.** (A) Two nucleotides of the DNA double helix, showing their antiparallel orientation, two H-bonds between A and T and three H-bonds between G and C; (B) the DNA double helix showing the major and minor grooves as well as the diameter of the molecule; (C) the convention of 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′-GCGCGCGCGCGCGC-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

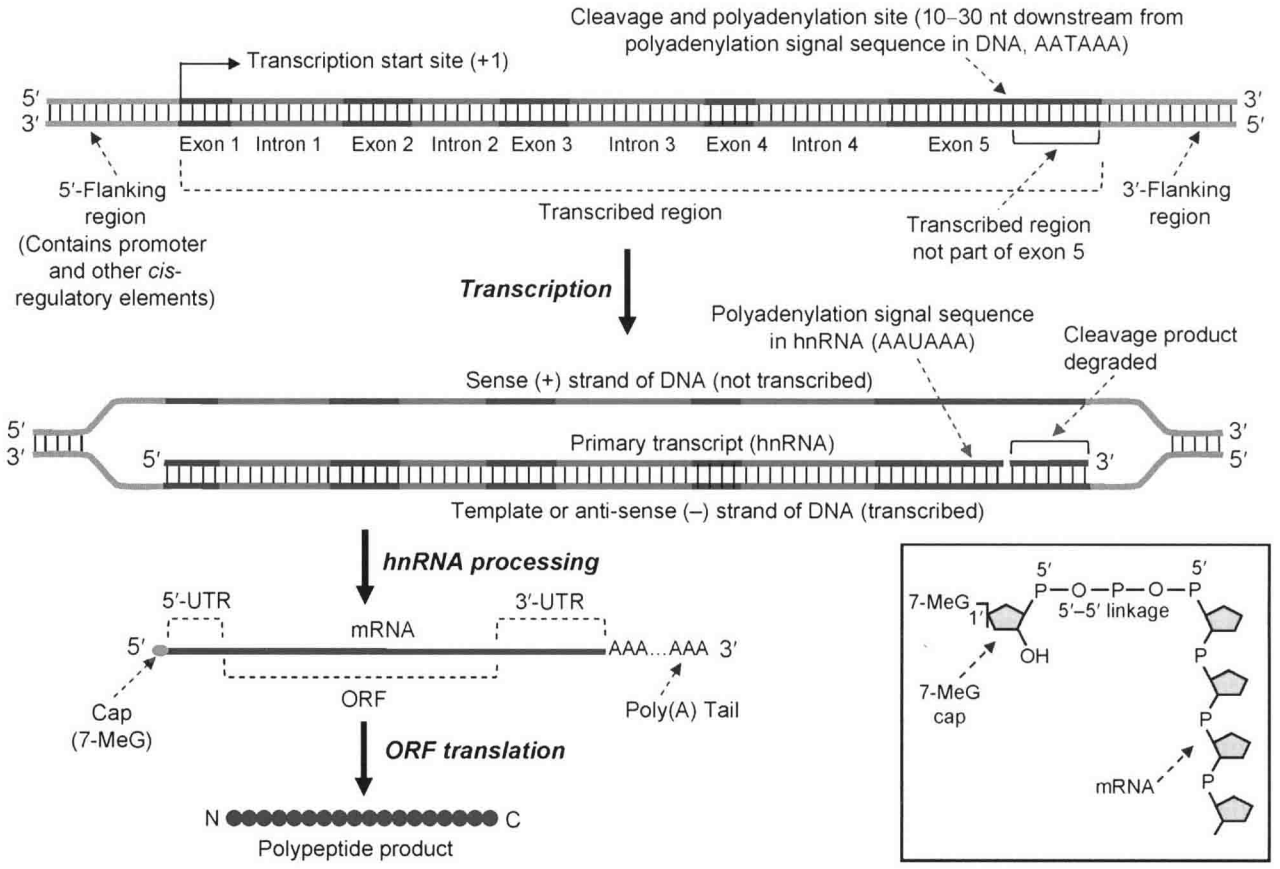


FIGURE 1.2 **Gene–hnRNA–mRNA–protein relationship.** Exon 1 is noncoding. Thus, the 5′-untranslated region (5′-UTR) is derived from exon 1, and the 3′-UTR is derived from the noncoding part of exon 5, which is the last and the longest exon. The sense strand of DNA has a "T" where the mRNA has "U"—for 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 **anti-sense** 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′-...NNN<u>AG</u>**gt**<u>a</u>nnn...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; <u>AG</u> are the last two bases of the preceding exon, and <u>a</u> is the base that immediately follows the splice donor site).

3′-splice site: 5′-...nnn<u>c</u>**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; <u>c</u> 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).

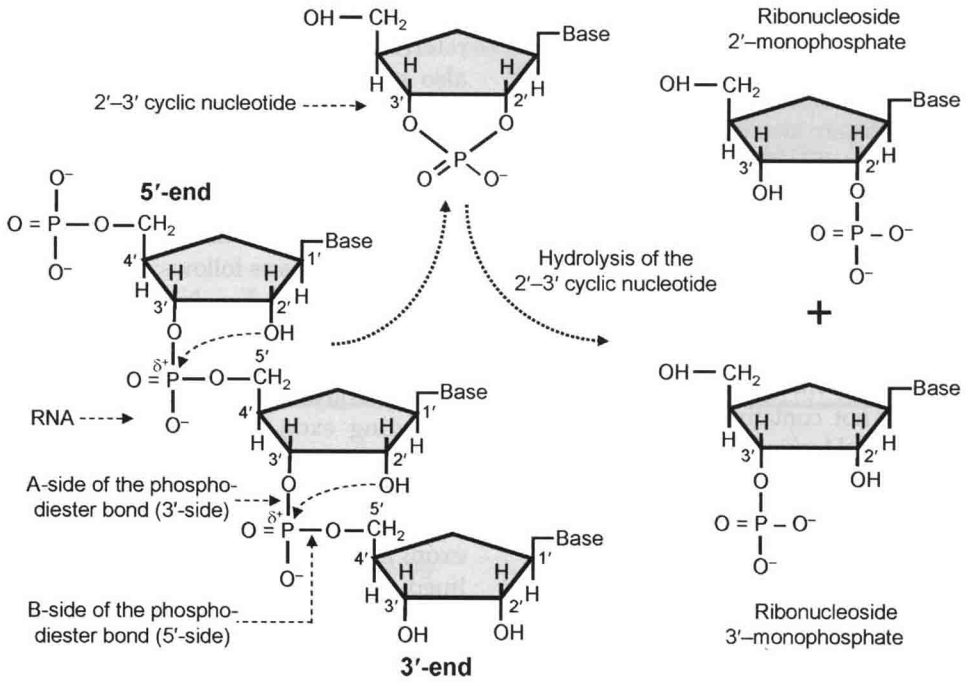


FIGURE 1.3 **Alkaline hydrolysis of RNA.** In an alkaline pH, the OH^- can abstract the H from the 2′-OH of ribose, generating the nucleophile 2′–O^-, which carries out a nucleophilic attack on the δ^+ P of the phosphate. This results in the cleavage of the phosphodiester bond and the formation of 2′–3′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

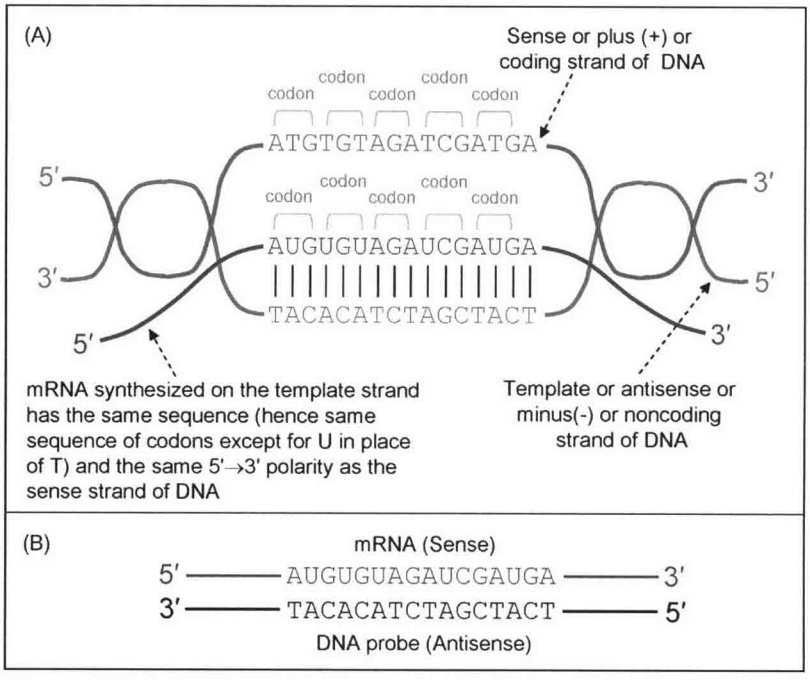


FIGURE 1.4 **Sense and antisense strands of DNA.** (A) The two strands of DNA have been drawn in different colors so that their respective 5′- and 3′-ends could be easily distinguished. The figure shows that mRNA and the sense strand have the same sequence (except for "U" in RNA and "T" in DNA) and the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