CPS260/BGT204.1 Algorithms in Computational Biology **August 28, 2003**

Lecture 2: Biological Background

Lecturer: Pankaj K. Agarwal *Scribe: Pankaj K. Agarwal*

2.1 Biomolecular Structures

The genetic information of all cells is encoded in the arrangement of nucleotides in their deoxyribonucleic acid (DNA). That information is first expressed via the formation of a related nucleic acid, ribonucleic acid (RNA), which in turn participates in the production of specific proteins. Genetic continuity from one cell generation to the next requires DNA replication, the process by which parental DNA molecules are duplicated. In all cellular organisms, gene expression requires that DNA be copied into RNA (transcription) followed by translation of RNA into proteins. DNA is transcribed into several kinds of RNA one of which messenger RNA (mRNA) encodes protein structures. This information relationship between the various biomolecular structures is shown in Figure 2.1.

There also exist a class of viruses called *retroviruses* that use an enzyme called *reverse transcriptase* to produce DNA from RNA. These viruses store their genetic information as RNA and then, after infecting a host cell, use reverse transcriptase to produce DNA that the host cell will later transcribe. The reverse transcriptase actually is a boon for biotechnology: it enables us to isolate mRNA and produce cDNA (details of what this means will covered later in the course). Although the DNA structure is relatively stable, it allows the coded information to change on rare occasions, called *mutations*, which provides genetic variation.

2.1.1 DNA

All cellular DNA consist of two polynucleotide side-by-side chains, called *strands*, wound around a common axis in the shape of a double helix: the DNA double helix. The polymer is composed of four different but related monomer units. Each monomer unit, known as a *nucleotide*, contains a *phosphate* group, a *deoxyribose sugar*, and a distinctive heterocyclic nitrogenous base, either a purine (a pair of fused rings)—*adenosine* (A) or *guanine* (G)—or a pyrimidine (a single ring)—*cytosine* (C) or *thymine* (T). See Figure 2.2. The backbone of each strand is a phosphate-deoxyribose sugar polymer. The sugar phosphate bonds in the backbone are called *phosphodiester* bonds. When two nucleotides polymerize to form nucleic acids, the hydroxyl (OH) group attached to the $3'$ C-atom of sugar of one of nucleotides forms an ester bond with the phosphate group of the other nucleotide, eliminating a water molecule; see Figure 2.3. Therefore, one end of a phosphate group of a nucleotide in a strand is connected to the $5'$ C-atom of one deoxyribose sugar and the other end is connected to the $3'$ C-atom of sugar of the next nucleotide. This induces a $5'$-to-$3'$ orientation of a backbone, which is an extremely important property of DNA molecules. The backbones of two strands in a DNA molecule are antiparallel.

The two strands of a DNA are held together by weak hydrogen bonds (H-bonds) between the bases of each

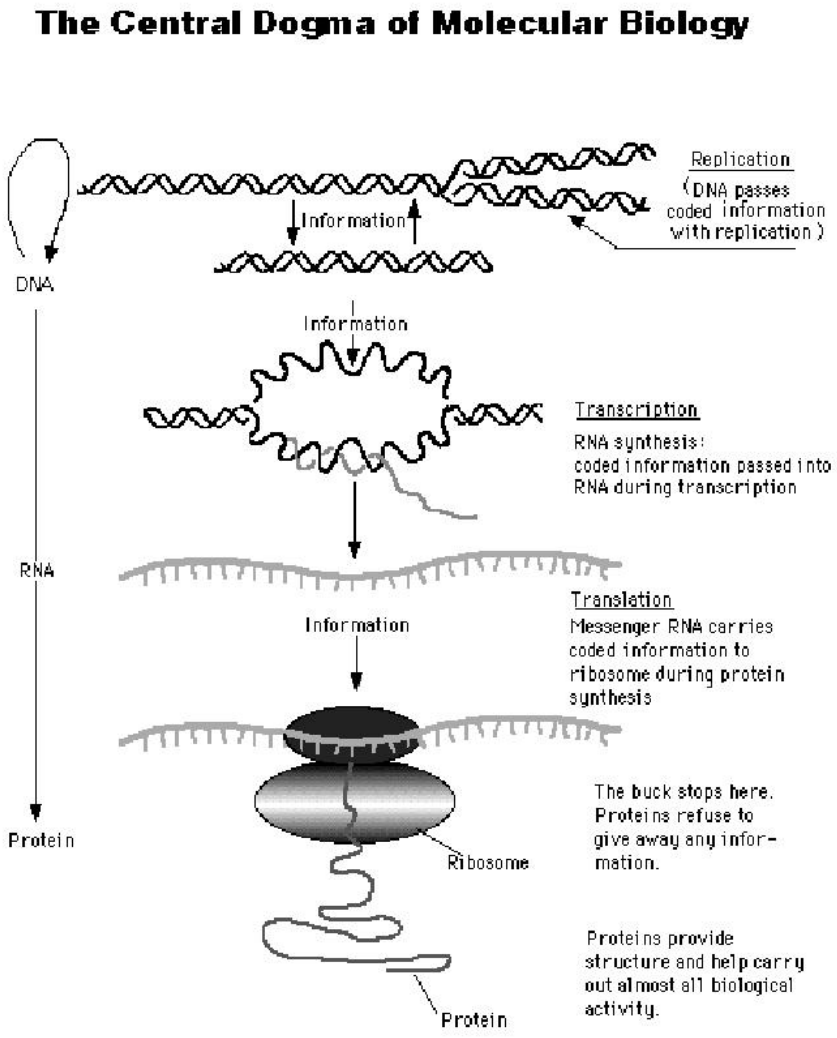


Figure 2.1: Informational relationships between DNA,RNA and proteins [4]

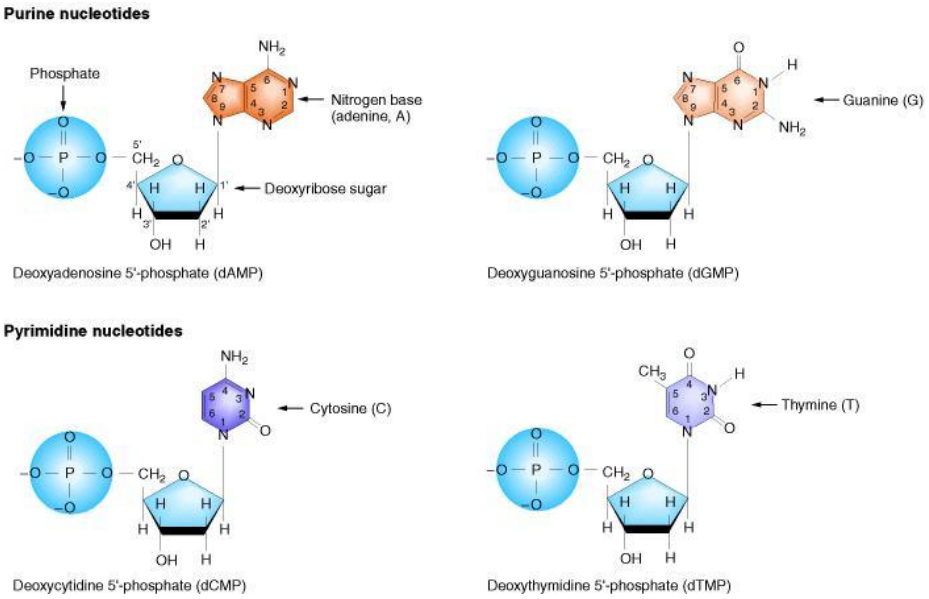


Figure 2.2: The nucleotide [1] [5].

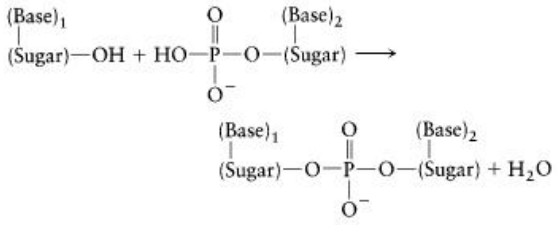


Figure 2.3: Two nucleotides combining to form a nucleic acid.

strand. Each base, attached to the $1'$ C-atom of deoxyribose, interacts with one base of the other strand. Two kinds of base pairs, often referred to as *complementary* (also known as *Watson-Crick*) base pairs, predominate in most DNAs: A with T and G with C. The A–T base pairs have two hydrogen bonds and hence are weaker than the G–C base pairs which have three hydrogen bonds. Hydrophobic and van der Waals interactions between adjacent base pairs also help stablize the DNA structure. Although DNA strands can form both left-hand and right-hand helices, natural DNA is right handed. The stacked base pairs are spaced about 0.34nm apart along the axis of the helix; the helix makes a complete turn every 3.4nm; so there are 10 base pairs per turn. This is referred to as *B-form* DNA. Other forms of DNA include A-form (right handed, 11 base pairs per turn) and Z-form (short molecule, left handed). There are circular DNA molecules (e.g. prokaryotic DNA, mitochondrial DNA), and there also exist triple stranded DNAs in test-tubes. Figure 2.4 shows different representations of the double helix, and Figure 2.5 show the replication of a DNA molecule.

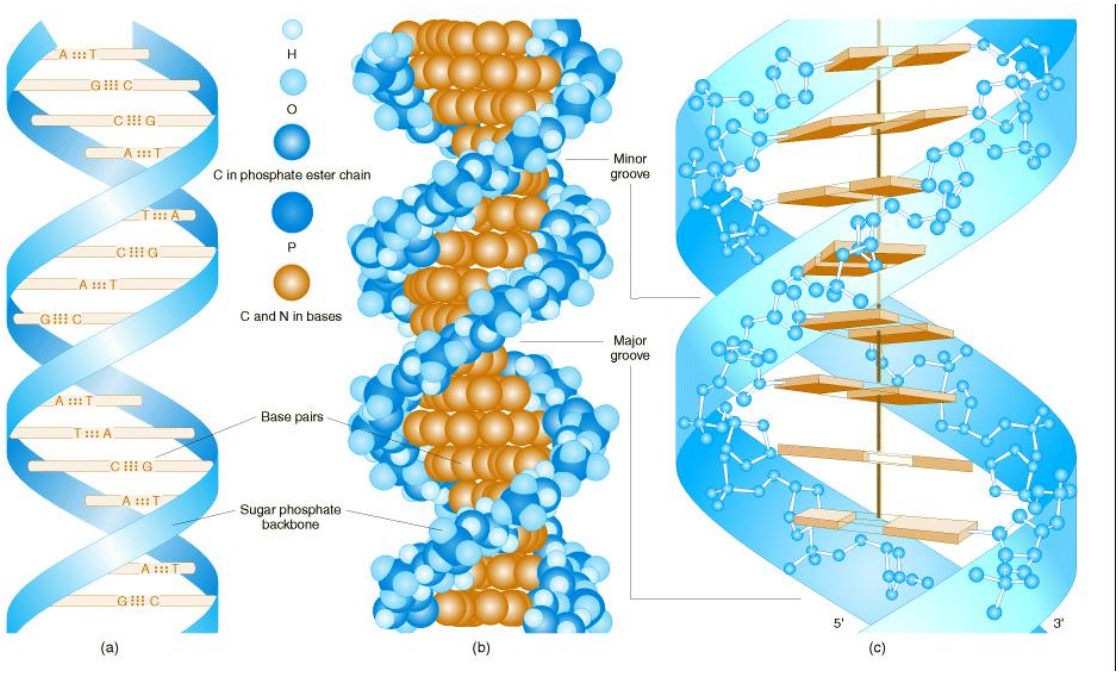


Figure 2.4: Different representations of DNA double-helix [5].

2.1.2 Genes

A gene is a region of DNA capable of being transcribed to produce a RNA, which is subsequently translated to a protein. Genes make up about 1% of the total DNA of our genome. The average human gene consists of 3Kbp, but sizes vary a lot, with the largest known human gene being *dystrphin* at 2.4Mbp.

One end of each gene has a regulatory region that enables it to receive and respond to signals in order it to be transcribed at the right moment. Not all genes are transcribed in each cell. In many eukaryotic DNA (including human), the coding portion of a gene, called *exons*, is interrupted by intervening sequences called *introns*. Both exons and introns are transcribed into pre-RNA, but then introns are removed by a splicing

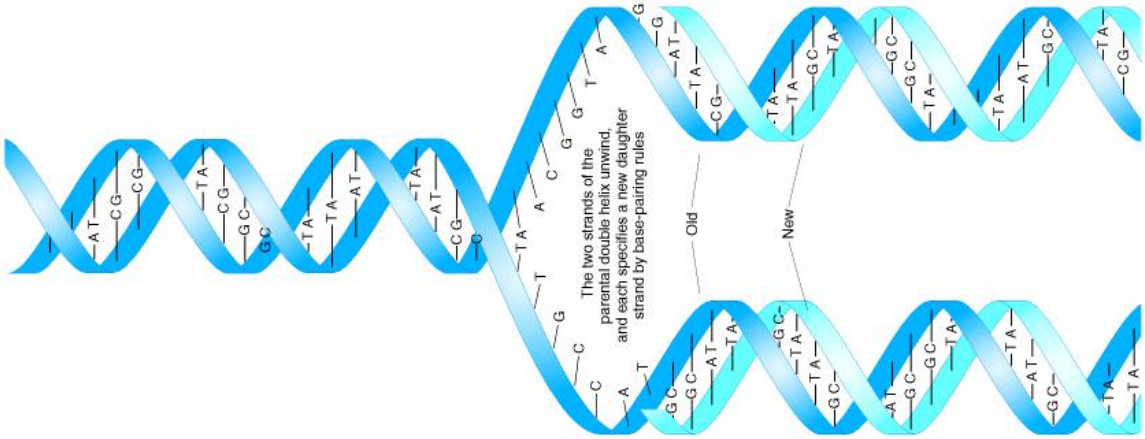


Figure 2.5: Replication of a DNA molecule [5].

machinery to produce a mRNA, which is then converted to a protein.

2.1.3 RNA

RNA has ribose sugar in its nucleotides—rather than deoxyribose as in DNA; see Figure 2.7. RNA nucleotides carry the bases adenine, guanine, cytosine, and uracil (U) instead of thymine. Analogous to DNA, RNA has a phosphate-sugar backbone, but it is a single stranded chain and thus has a much more complex 3-dimensional shape than DNA. Several types of RNA occur in all cells:

Informational RNA: Informational RNAs are intermediate tarnscripts of DNA. A DNA molecule is first transcribed to primary RNA (pre-mRNA) and then to messenger RNA (mRNA).

Transfer RNA (tRNA): Acts as transporter that brings amino acids to mRNA during the protein synthesis. Each type of amino acid has its own type of tRNA, which binds it and carries it to the growing polypeptide chain.

Ribosomal RNA (rRNA): Components of ribosomes, which are macromolecuar assemblies that coordinate the conversion of mRNA to proteins.

Small nuclear RNA (snRNA): Involved in splicing of pre-mRNA into mRNA in the eukaryotic cell nucleus.

Small cytoplasmic RNA (scRNA): Involved in protein trafficking within eukaryotic cells.

2.1.4 Proteins

A protein consists of one or more *polypeptides* chains, each consisting of a long unbranched polymer of *amino acids*. All polypeptides, whether from viruses or humans, are constructed from a repertoire of only 20

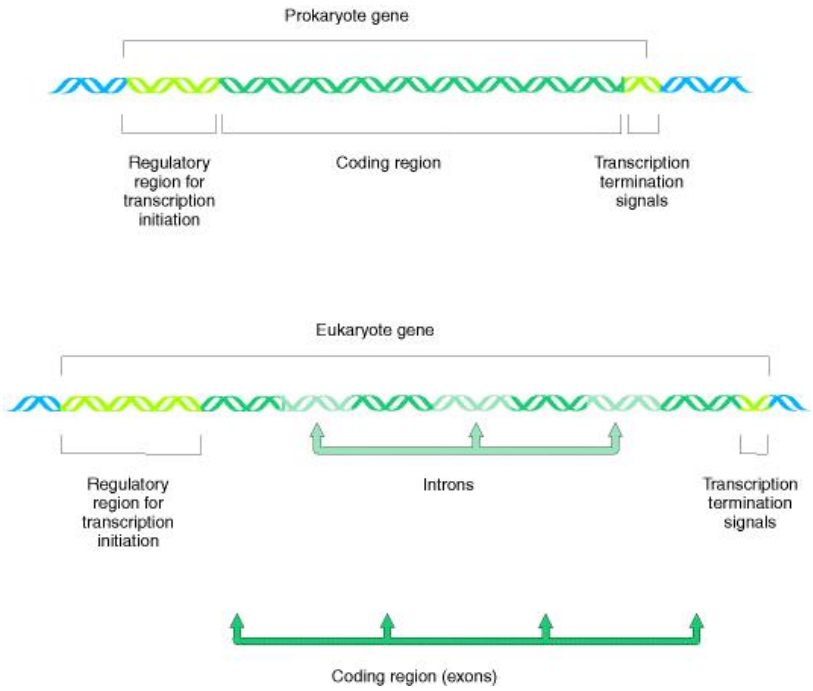
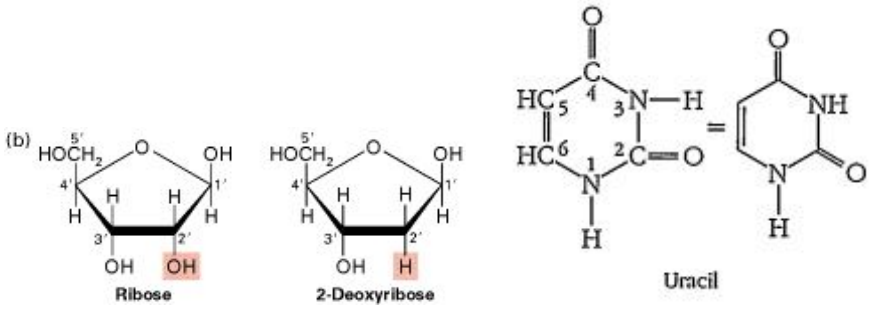
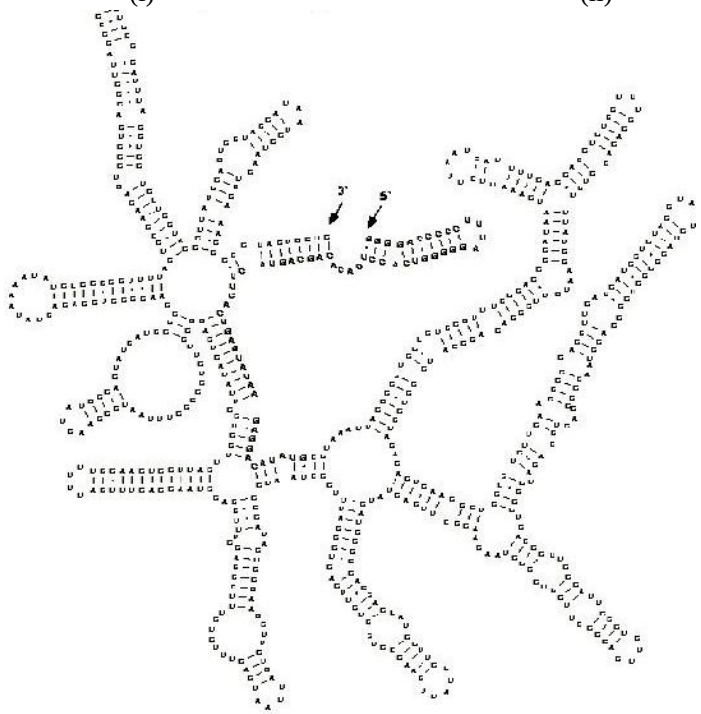


Figure 2.6: Structure of genes in human genome [5].



(i) (ii)



(iii)

Figure 2.7: (i) Ribose vs. deoxyribose (ii) Uracil, and (iii) secondary structure of a tRNA.

different amino acids. The general structure of an amino acid is shown in Figure 2.8 (i). The *side chain*, R (reactive group), depends on the amino acid. Figure 2.9 shows different amino acids.

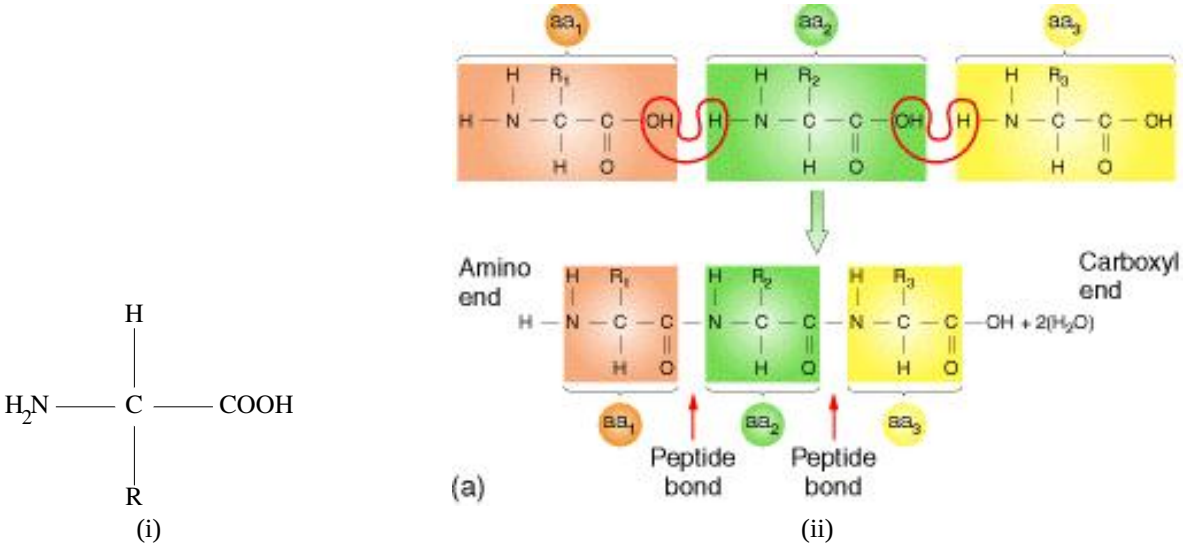


Figure 2.8: (i) An generic amino acid, (ii) A polypeptide chain [5].

Each protein possesses a unique amino acid order along its polypeptide backbone. This defined order of amino acids is called the *primary* structure. Each polypeptide chain has one amino (NH_2) end and one carboxyl (COOH) end, as shown in Figure 2.8 (ii). Several different types of forces acting between the atoms cause the protein to fold in a specific shape. The chains can fold into regularly repeating structures called *secondary* structures; α-helices and β-strands are the most common secondary structures. The *tertiary* structure of a protein is produced by the folding of secondary structures. Proteins come in all forms and shapes. Virtually all enzymes and regulatory proteins have *globular*—compactly folded—three dimensional structures. Some proteins such as fibrous proteins have linear structure. If a number of folded polypeptides form a complex unit, we can speak of the unit's *quaternary* structure.

2.2 Transcription: DNA to mRNA

Figure 2.10 provides an overview of RNA and protein synthesis.

The RNA synthesis relies on complementary base pairing between DNA and RNA nucleotides. The two strands of the DNA double helix separate locally, and one of them acts as a *template* for RNA synthesis. Catalyzed by RNA polymerase, free ribonucleotides in the cell align with the DNA template—A (resp. G, C, U) of RNA nucleotides aligns with T (resp. C, G, A) of DNA nucleotides. RNA always grows in the $5'$-$3'$ direction—RNA nucleotides are added at a $3'$ growing tip, and the DNA template strand must be oriented $3'$-$5'$. The nontemplate strand of DNA is oriented $5'$-$3'$ and has the same sequence as the growing RNA (except U being replaced with T), and thus a gene DNA sequence is referred to the sequence of this nontemplate strand. See Figure 2.11. Transcription is performed in three stages:

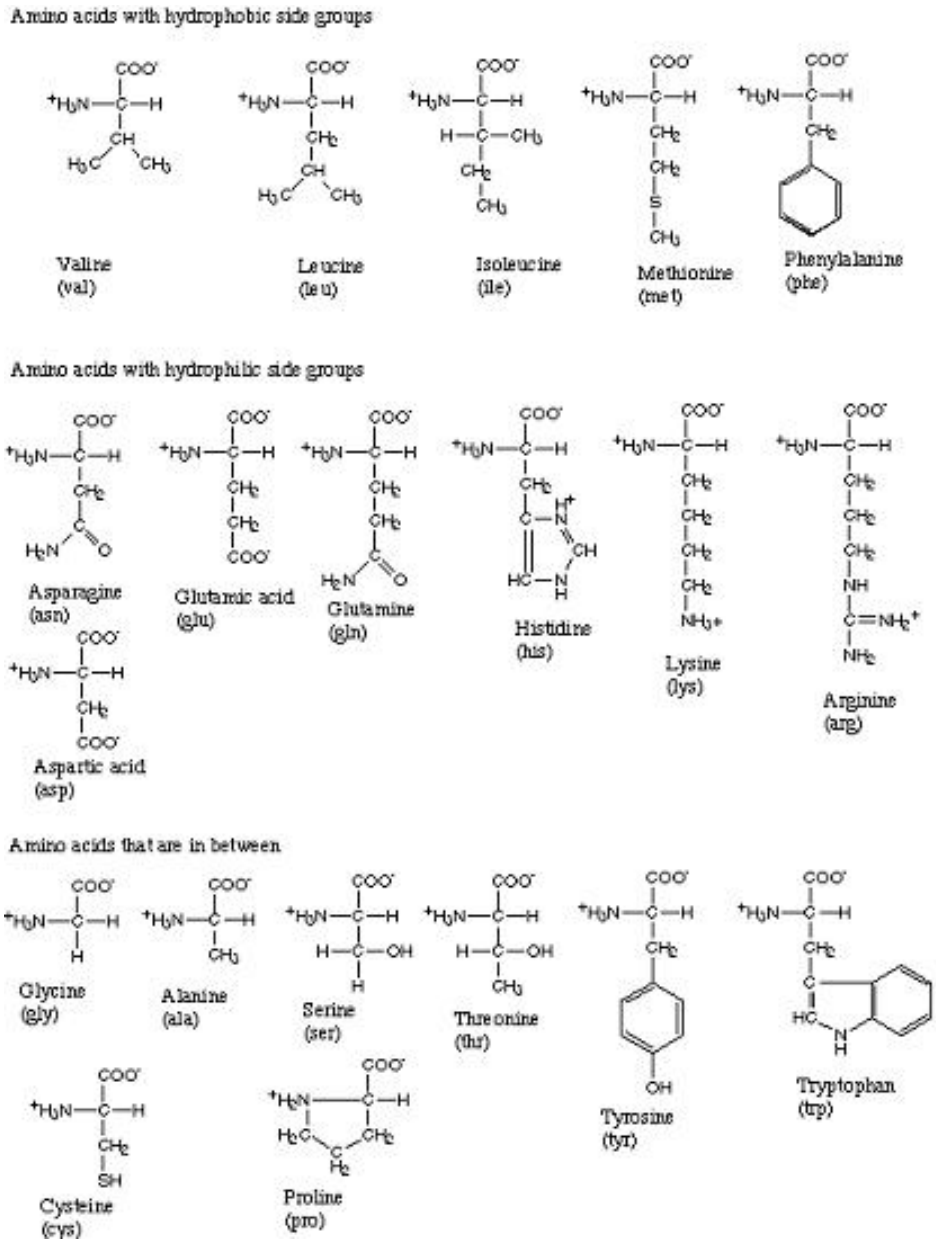


Figure 2.9: Different types of amino acids.

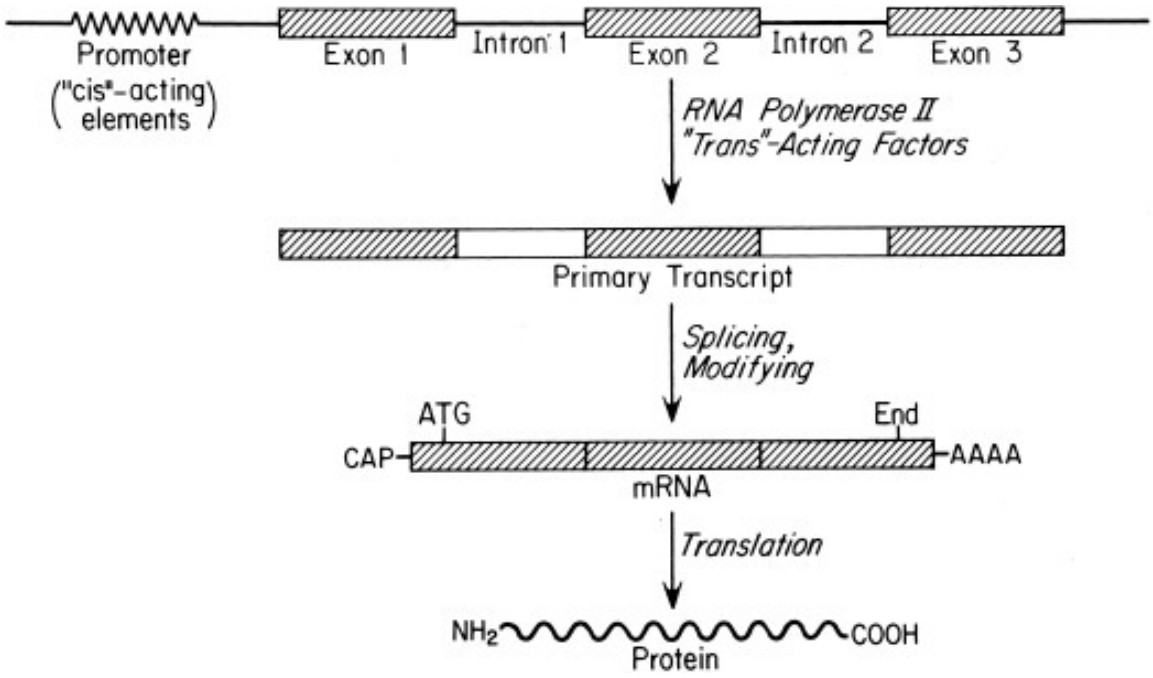


Figure 2.10: RNA processing overview

Initiation: RNA polymerase binds to a "specific" DNA sequence called *promoter*, which is part of the regulatory region adjacent to the encoding region of gene, and initiates transcription. After the initial binding, RNA polymerase unwinds the DNA and begins the synthesis of RNA.

Elongation: RNA polymerase moves along the DNA, maintaining a transcription "bubble" to expose the template strand and catalyzing the $3'$ elongation of the RNA strand.

Termination: When RNA polymerase recognizes specific nucleotide sequence in the DNA that act as signals for chain termination, the RNA strand and the polymerase are released from the template. The terminal sequence consists of about 40bp ending in GC-rich stretch followed by a run of six or more A's on the template strand.

Transcription of eukaryotic DNA. Recall that genes in many eukaryotic cells (including human cells) have introns and exons, of which only exons carry the functional information. Thus in the first step of transcription, a pre-mRNA is synthesized, which contains both introns and exons. Then a splicing step removes introns, leaving a contiguous sequence of exons and converting pre-mRNA into a mature mRNA. Almost all exon-intron junctions have specific sequences — GU at the $5'$ splice site of the intron (in pre-RNA) and AG at the $3'$ splice-site. The sequences are recognized by small nuclear ribonucleo protein particles (snRNPs), which catalyze the splicing process. Figure 2.12 illustrates the splicing step in detail. These steps along with the splicing step are shown in Figure 2.13.

Before the crucial splicing step, a few other processes occur during the transcripton of the DNA of a eu-

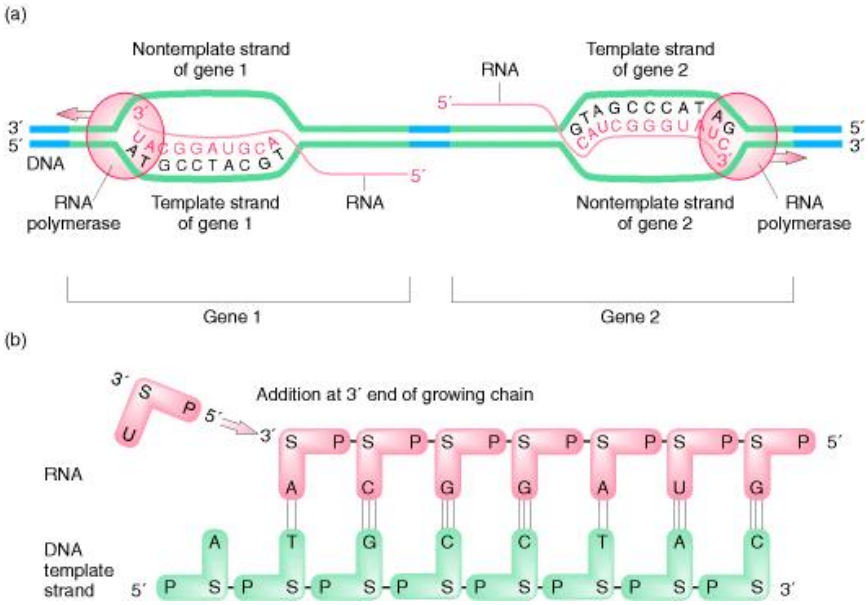


Figure 2.11: Initial transcription of DNA [5].

karyotic cell. First a *cap* consisting of a 7-methylguanosine ($\mathrm{m}^7\mathrm{Gppp}$) residue is added to the $5'$ end of the transcript. Then an AAUAAA sequence near the $3'$ end is recognized by an enzyme that cuts off the end of the RNA approximately 20 bases farther down (this step is called *cleavage* step). At this time a stretch of 150 to 200 A's are added at the cut $3'$ end, this final step is called *polyadenylation*.

2.3 Translation: mRNA to Protein

The mRNA is transported out from the nucleus to cytoplasm where it encoutners ribsomes, protein-building factories. A number of ribsomes move along the mRNA, each starting at the $5'$ end and proceedings all teh way to its $3'$ end. As a ribosome moves along, it "reads" the sequence of the mRNA, three nucleotides at a time. Each triple, called *codon*, stands for a specific amino acid. There are $4^3 = 64$ codons but only 20 amino acids, so many codons correspond to the same amino acid; see Figure 2.14.

Amino acids are added to a tRNA, each tRNA recognizes a specific amino acid; see Figure 2.15. Amino-acid bound tRNAs, called *aminoacyl-tRNA*, and mRNAS meet at ribsomes. Ribosomes consist of a large and a small subunit, each composed of many rRNAs and proteins. The mRNA binds to the small subunit, and the tRNA binds to two sites that overlap the subunits. The A-site is the entry site for an aminoacyl-tRNA and the P-site is the exit site. Each new amino acid is added by the transfer of the growing chain to the new aminoacyl-tRNA, forming a new peptide bond. The tRNA that released its amino acid exits through the P-site, and ribosome moves one codon farther along the mRNA. See Figure 2.15. The growing end of peptide chain is the carboxyl (COOH) end, and the free end if the amino (NH_2) end. Hence the amino end corresponds to the $5'$ end of the mRNA and the carboxyl end corresponds to the $3'$ the end. The translation also occurs in three main phases.

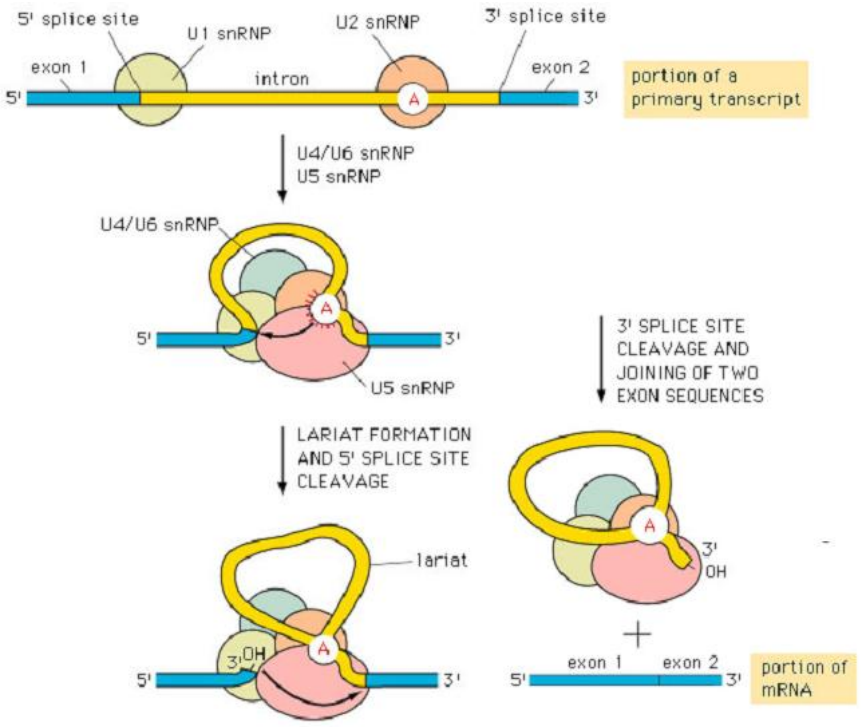


Figure 2.12: Details of the splicing step.

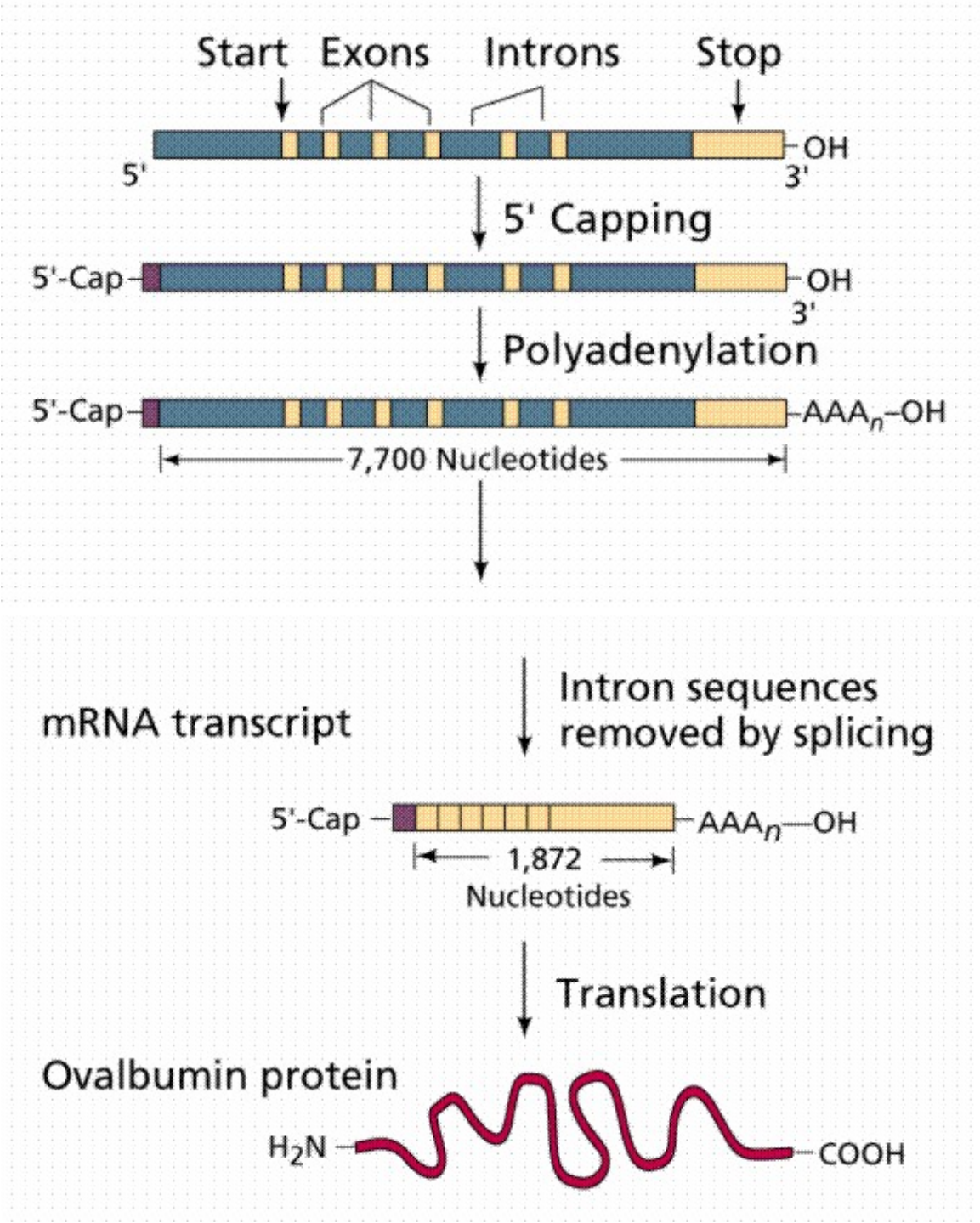


Figure 2.13: Transcription of a eukaryotic DNA [2].

SECOND POSITION

FIRST POSITION	U	C	A	G	THIRD POSITION
U	phenyl-alanine	serine	tyrosine	cysteine	U
					C
	leucine		stop	stop	A
			stop	tryptophan	G
C	leucine	proline	histidine	arginine	U
					C
			glutamine		A
					G
A	isoleucine	threonine	asparagine	serine	U
					C
	* methionine		lysine	arginine	A
					G
G	valine	alanine	aspartic acid	glycine	U
					C
			glutamic acid		A
					G

* and start

Figure 2.14: Correspondence between codons and amino acids.

Initiation: In eukaryotic mRNA, a ribosome carrying an initiator tRNA attaches under guidance of the $5'$-cap, moves along the mRNA and initiates the translation at the first "appropriate" AUG (methionine) codon, the universal starting codon, it encounters.

Elongation: Elongation is assisted by several proteins, called *elongation factors*, that guide the binding and movement of the tRNA and ribosome.

Termination: When ribosome encounters a stop codon (UAG, UGA, UAA), the release factors (proteins) bind to the A-site of the ribosome, athe polypeptide is then released through the P-site, and the ribosome dissociates into two subunits ending the translation.

Figure 2.8 depicts how proteins are assembled during translation.

Bibliographic Notes.

Sections 2.1.1–2.1.3 are taken from [5] and [6]. See Branden and Tooze and Creighton for details on proteins. Sections 2.2 and 2.3 are taken from [5]. See [2] for details on ribosomes. Figures are taken from various sources, as cited in their captions. They cannot be reproduced or distributed without permission from the original sources.

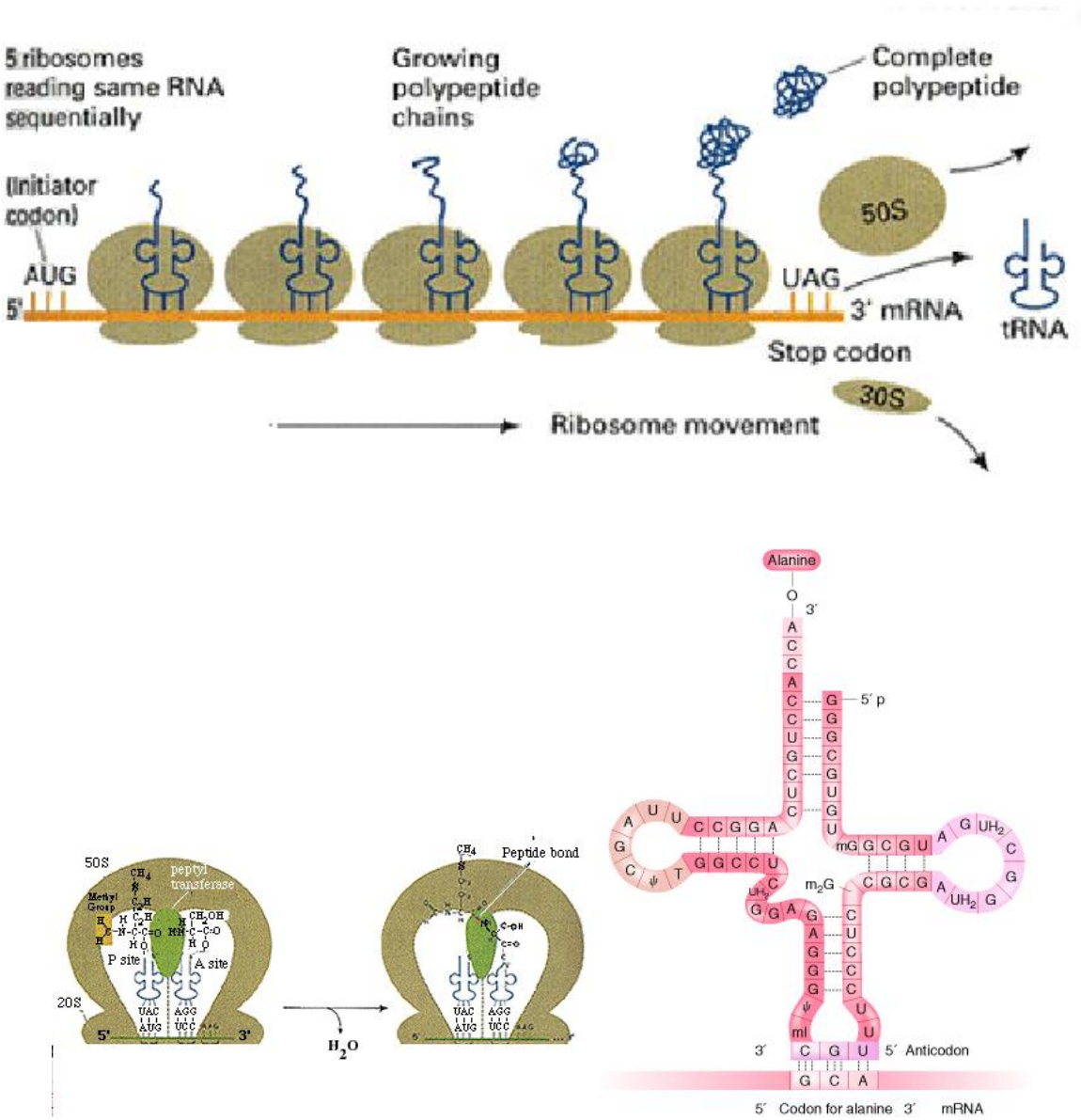


Figure 2.15: Protein Synthesis [2] [5].

References

[1] http://gened.emc.maricopa.edu/bio/bio181/BIOBK/BioBookDNAMOLGEN.html. *DNA and Molecular Genetics*, 2001. 2-3

[2] http://ntri.tamuk.edu/cell/ribosomes.html. *Ribosome Structure and Function*, 2001. 2-13, 2-14, 2-15

[3] http://www.blc.arizona.edu/Molecular_Graphics/DNA_Structure/DNA_Tutorial.HTML. *Introduction to DNA structure*, 2001.

[4] http://www.postmodern.com/ jka/rnaworld/nfrna/nf-rnadefed.html. *The Central Dogma of Molecular Biology*, 2001. 2-2

[5] A. Griffiths, W. Gelbart, J. H. Miller, and R. Lewontin. *Modern Genetic Analysis*, W. H. Freeman, 1999. 2-3, 2-4, 2-5, 2-6, 2-8, 2-11, 2-14, 2-15

[6] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell. *Molecular Cell Biology* (4th ed.), W. H. Freeman, New York, 1999.

2-14