2.3 RNA

As with DNA, RNA is a macromolecule built from repeating subunits. Deoxyribonucleic acid has one less oxygen atom (deoxy) in the ribose sugar than Ribonucleic Acid (RNA). Specifically, RNA has a hydroxyl (OH) attached to the 2′ carbon on the ribose sugar and DNA has a hydrogen atom (H) attached at that site on the sugar. In RNA, the pyrimidine base uracil (U) replaces thymine. As with the pyrimidine thymine, uracil complements adenine (A). These subunits are comprised of a nitrogenous base, a ribose sugar, and a phosphate group generically denoted NTP (note dNTP was used for DNA) for ribonucleotide triphospate. The NTPs are ATP, CTP, GTP, and UTP.

Ribonucleic acid usually occurs as a single strand. The single-stranded structure is less stable than the double-stranded DNA structure. For analysis and manipulation, RNA is often copied into a complementary strand of DNA that can be paired into a double strand.

2.4 How DNA Codes for Protein

2.4.1 Transcription

Transcription is the process of creating an mRNA strand that contains genetic information from the DNA. DNA information is transcribed to a single-stranded RNA messenger that delivers the genetic information to a ribosome. This is illustrated in Fig. 2.6 for prokaryotic cells. First, proteins known as helicases unwind a portion of double-stranded DNA. The entire chromosome is not unwound—the hydrogen bonds connecting base pairs are broken locally. The region of interest on the DNA is often preceded by a promoter section of DNA. Specific proteins bind to the promoter region and "promote" attachment of an enzyme known as RNA polymerase. Synthesis of a complementary RNA strand begins near the promoter position and moves along the DNA strand from the 3′ to the 5′ end. Because the RNA is complementary to the DNA, it is synthesized from the 5′ to the 3′ end. In Fig. 2.6 and as a convention for this book, we will represent double-stranded DNA with the top strand going from 5′ on the left to 3′ on the right. Therefore the RNA polymerase is attached to the lower strand and is moving from left to right (3′ to 5′ on the DNA). There are free nucleotides available in solution as ATP, CTP, GTP, and UTP and these polymerize to their complements catalyzed by the RNA polymerase.

As shown in Fig. 2.6, if 5′-AAT-3′ is the top strand, then 3′-TTA-5′ is the bottom strand and the RNA should be 5′-AAU-3′. Recall that uracil in RNA replaces thymine in DNA. This is the mRNA that communicates the genetic information to the ribosomes and it is the same as the "sense" 5′-to-3′ strand (the top strand in the figure with 5′ at the left side) with uracil replacing thymine. The "nonsense" or "antisense" strand is the bottom strand in the figure and is also known as the template strand because the polymerase enzyme actually moves along that strand (from the 3′ to the 5′ end).

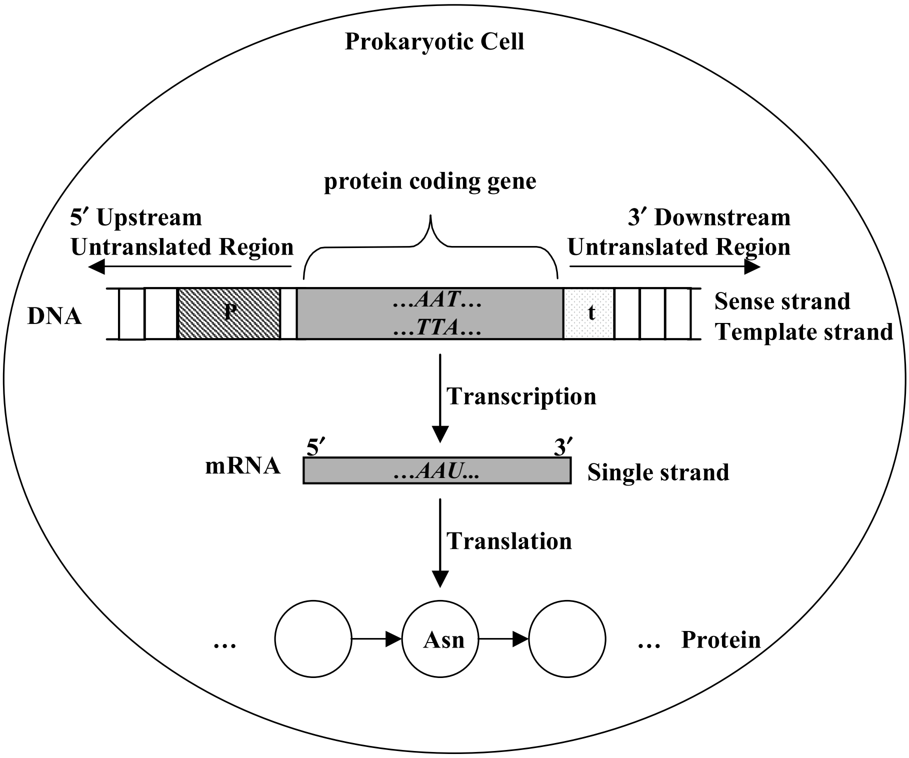
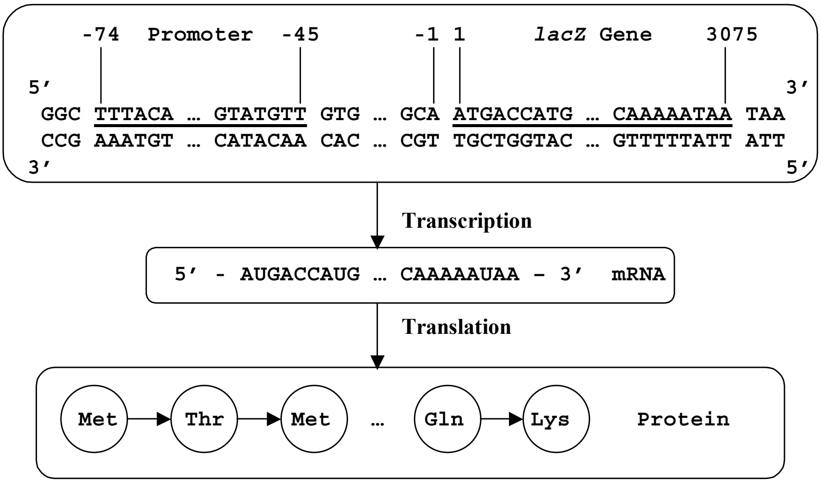


Fig. 2.6. Transcription of a gene requires the conversion of DNA into mRNA by an enzyme that attaches to the template strand near a promoter (P) and moves toward the terminator (t). Groups of three bases (codons) in the mRNA code for specific amino acids that are assembled into chains of amino acids on ribosomes. The process of mRNA directing the formation of amino acid chains (proteins) is known as translation. The codon AAU codes for the amino acid asparagines (Asn).

Some other terms are easily defined with Fig. 2.6. The top strand region to the left is known as the 5′ upstream. The region on the right is the 3′ downstream. By orienting the strands with the RNA polymerase running left to right, nucleic acids of interest (DNA and mRNA) usually run 5′ to 3′ left to right. It is common to only list the top 5′ to 3′ strand since the second strand can be generated by complementing the bases. The sections of DNA that code mRNA are known as protein coding regions. Transcription begins near a promoter site and ends at a terminator site. In prokaryotes, genes tend to be represented by contiguous bases in the DNA, with promoter and other transcription factors nearby. There are short three base DNA sequences that code for start (ATG) and stop (TGA, TAG, TAA). Unfortunately, other factors also mediate this process, making the start and stop codes useful markers but not sufficient to identify genetic regions.

A specific example of transcription in prokaryotic cells is given in Fig. 2.7. The 3,075-base *lacZ* gene in *E. coli* is shown to transcribe into a 1,023-amino acid protein. The 30-base promoter region includes two binding sites (TATGTT and TTTACA). Some of the data available at the National Center for Biotechnology Information (NCBI) online at http://www.ncbi.nlm.nih.gov/ are also listed for the DNA sequence of the promoter region and *lacZ* gene. Note that the NCBI data are for the opposite strand of the DNA—it was labeled online as the "complement." In order to arrange the DNA sequence in Fig. 2.7(a) consistent with the sense strand on top going left to right from 5′ to 3′, the NCBI data must be reversed and complemented. Simply reversing the NCBI data provides the bottom or template strand for the *lacZ* gene.


-74 Promoter -45
-1 1
lacZ Gene 3075
5′
3′
GGC TTTACA … GTATGTT GTG … GCA ATGACCATG … CAAAAATAA TAA
CCG AAATGT … CATACAA CAC … CGT TGCTGGTAC … GTTTTTATT ATT
3′
5′
Transcription
5′ - AUGACCAUG … CAAAAAUAA - 3′ mRNA
Translation
Met
Thr
Met
…
Gln
Lys
Protein

(a)

```
5581 AGTACATAAT GGATTTCCTT ACGCGAAATA CGGGCAGACA TGGCCTGCCC GGTTATTATT
5641 ATTTTTGACA CCAGACCAAC TGGTAATGGT AGCGACCGGC GCTCAGCTGG AATTCCGCCG
     …
8641 GTTGGGTAAC GCCAGGGTTT TCCCAGTCAC GACGTTGTAA AACGACGGCC AGTGAATCCG
8701 TAATCATGGT CATAGCTGTT TCCTGTGTGA AATTGTTATC CGCTCACAAT TCCACACAAC
8761 ATACGAGCCG GAAGCATAAA GTGTAAAGCC TGGGGTGCCT AATGAGTGAG CTAACTCACA
```

(b)

Fig. 2.7. Transcription and translation demonstrated using genes in the lactose metabolism operon of *E. coli.* The sequence data for this example, partially repeated in (b), are available online at the National Center for Biotechnology Information (NCBI) web site. (a) Specific example of transcription and translation for the *lacZ* gene related to lactose metabolism in the bacterium *E. coli.* The transcribed strand (sense strand) is shown on top with 5′ at the left end. (b) Data from NCBI AE000141 *E.* coli K12 MG1655 for the *lacZ* gene (complement of the 3075 bases 8713 to 5639) and its promoter (complement of the 30 bases 8787 to 8758). There are 44 bases between the promoter and the start codon.

Transcription in eukaryotic cells is more complex and is outlined in Fig. 2.8. In eukaryotic DNA, noncoding regions (introns) interrupt protein-coding regions (exons). Introns range in size from 40 to 10,000 bases. Eukarya have very complex promoter logic, often requiring multiple sites and multiple proteins to promote transcription. Sometimes large protein complexes span several sites on the DNA. Collectively these regulatory proteins are referred to as transcription factors. A first transcript or principal transcript of the DNA strand is made that includes RNA that complements both the exons and the introns. In addition to the bases from the DNA template, there are also bases appended to the ends of the principal transcript. At the 5′ end, a G base is appended and is known as the guanine cap. At the 3′ end, a string of up to 200 adenine bases is appended and is known as the poly(A) tail or polyadenylation. A second RNA strand known as the functional transcript is made by splicing exons together between the G-cap and the poly(A) tail. Occasionally some exons are omitted during splicing and an alternative protein is coded in the functional transcript. The functional transcript then migrates outward from the nucleus toward ribosomes in the cytoplasm.

The information stored in the DNA sequence of a gene is transcribed into a message of single-stranded RNA. On a ribosome in the cytoplasm, the mRNA is translated one codon (three bases) at a time into one of 20 amino acids that are also sometimes referred to as residues. The amino acids designated by the codons are chained together until a full-length protein is formed (see Fig. 2.9). Peptides are short chains of amino acids less than 40 residues long. Most functional proteins are longer than 40 residues. The amino acids have a modular structure built around a central carbon (Cα) flanked by a hydrogen (H) atom, an amino group (NH_2), a carboxyl group (COOH), and a side chain (R) that defines the specific amino acid of the 20 possible. Note that "R" in this case is used for "residue" and not to denote the specific amino acid arginine. Figure 2.10 is a chemical schematic of an amino acid structure. The amino acids are joined together by peptide bonds where the carboxyl group gives up an OH and the amino group donates an H. The bond between the carbon and the nitrogen atoms of the carboxyl and amino groups is known as a peptide bond. As with DNA, amino acid chains can be oriented using the N-terminal group (NH_2) or the C-terminal group (COOH) of the peptide backbone. Proteins are the workhorses of the cell, performing chemical (enzymatic) or structural functions. DNA and the environment control the quantity, timing, and selection of proteins expressed.

The ribosome uses the mRNA and another type of RNA called transfer RNA (tRNA) to construct proteins. As shown in Table 2.1, there are 64 (four-cubed; three-base sets pulled from four possible bases A, C, G, and T) possible codons that redundantly code for the 20 amino acids. An amino acid is attached to the 3′ end of a charged single strand of RNA (the tRNA) with a complementary codon (anticodon) available to bind to the RNA. The pairing of the mRNA with the appropriate series of tRNA collects amino acids on the ribosome so that the formation of peptide bonds can produce a protein.

In many cells there are processes that interfere with the production of protein. In some cases the mRNA is consumed before translation. In eukaryotic cells, the mRNA may not successfully leave the nucleus to reach a ribosome. It is also possible to have very efficient DNA-to-mRNA to protein processes. Under some conditions, bacterial cells will have translation occurring at a ribosome on an mRNA that has not yet been fully transcribed! Cellular differences, biochemistry, and other factors affect translation efficiency. Because of these dependencies, an increase in mRNA transcription of a particular gene does not guarantee an increase in protein expression in the cell.

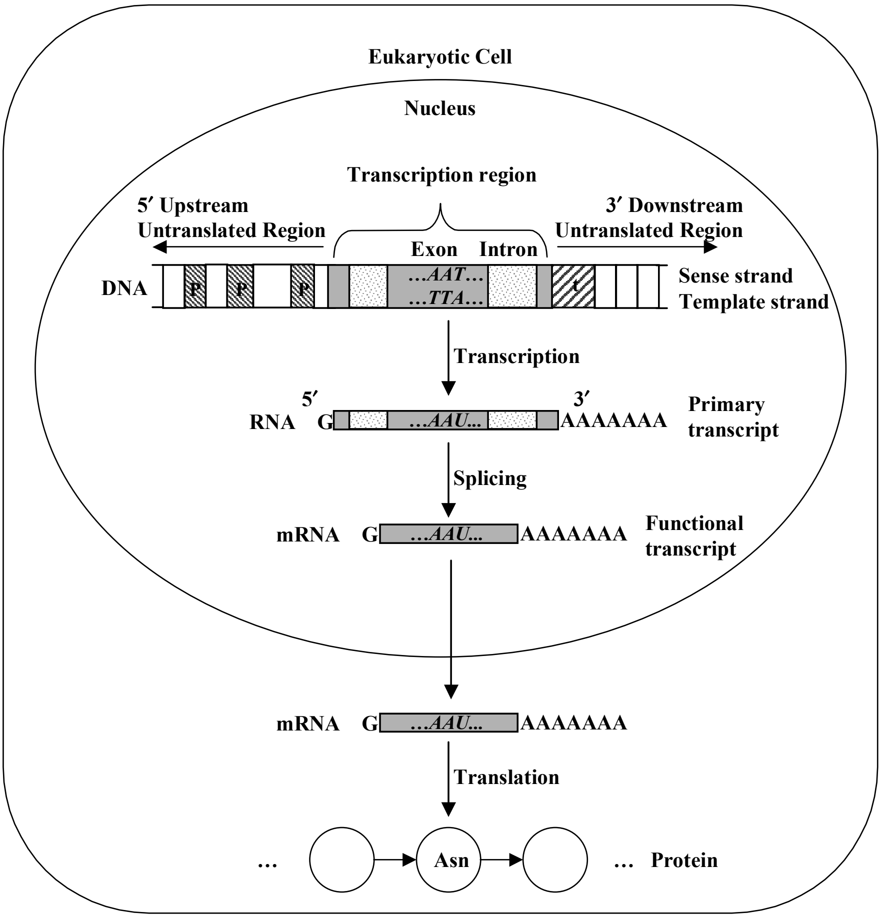


Fig. 2.8. Transcription of a gene in a eukaryotic organism requires the conversion of DNA into a primary RNA transcript as well as the splicing of exons (removal of introns) into the mRNA. The mRNA leaves the nucleus and is translated into protein on a ribosome. Promoters (P) and enhancers can be distributed in the untranslated regions near the gene. Eukaryotic organisms often have a regulatory DNA sequence in more locations and spread over more bases than prokaryotic cells.

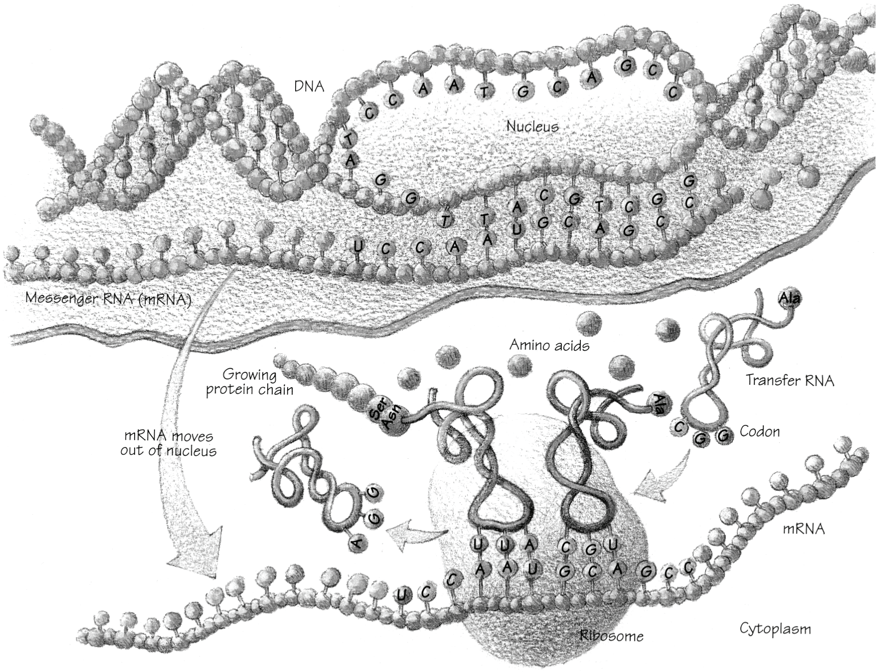


Fig. 2.9. Transcription of DNA into mRNA in the nucleus followed by translation of mRNA at the ribosome into a growing amino acid chain to form a protein. (From *To Know Ourselves*, U.S. Dept. of Energy Rept. PUB-773, July 1996 online at www.lbl.gov/Publications/TKO.)

As an example of working from DNA to protein, consider the double-stranded eukaryotic DNA shown in Fig. 2.11. The top strand is 5′ to 3′ from left to right. RNA is spliced into a primary transcript (b); coding regions (exons) are spliced together as the mRNA is formed (c); and translation of the mRNA into a protein completes the process (d).

Once translation is completed and the protein is away from the ribosome, the amino acid chain assumes a three-dimensional (3-D) shape. Protein structure and function are closely related. There are several levels for describing protein structure. The primary structure is the order of amino acids in the protein. By convention, the amino acids are listed from the amino end of the protein (N-terminal) to the carboxyl end (C-terminal). Recall that the amino acids are joined by peptide bonds. These bonds form with the removal of water (condensation or dehydration synthesis reaction). Protein secondary structures are common repeating structures found in many proteins known as the alpha helix and the beta-sheet. Alpha helices are the most common of the two and occur when hydrogen bonds form between the CO of one amino acid and the NH group of another amino acid four residues away. Beta-sheets or beta-pleated sheets are the other type of secondary structure. A tertiary protein structure is the full three-dimensional structure of the amino acid chain.