2 From simple genome to complex proteome

Trillions of cells that make up an adult human, originate from a unique ancestor - *the fertilized egg.* These cells largely contain the same genetic material, but are enormously diverse both structurally and functionally. The process by which different cell types emanate from a single cell is called *cell differentiation.* The cause of this differentiation is not variation in DNA but the contrasting interpretation of the identical DNA blueprint (*gene expression*). This interpretation of genetic information is tightly regulated, leading to the development of different cell types. This chapter introduces various aspects of information flow in biological cells, summarized as *the central dogma of molecular biology* (Figure 2.1). The dogma states that the flow of genetic information is DNA to RNA (*transcription, alternative splicing*) to protein (*translation*). Apart from a few exceptions, viz. retroviruses and prions, all biological cells conform to this rule.

2.1 Transcription

Transcription is a process in which one DNA strand is used as template to synthesize a complementary RNA. The DNA strand which serves as the template is called *template strand*, while the other DNA strand is as termed *coding strand.* Figure 2.2 illustrates the process. Since both DNA coding strand and RNA strand are complementary to the template strand, they have identical sequence except that the nucleotide `T`in the DNA coding strand is replaced by the nucleotide `U`in the RNA strand.

The process of transcription consists of four essential steps:

1. *Unwinding of the DNA double helix:* The DNA double helix needs to be unwinded so that it is accessible to the transcription machinery. In case of prokaryotes, the polymerase themselves direct the unwinding activity. However, for eukaryotes the enzyme *helicase* catalyzes the unwinding of the DNA double helix.

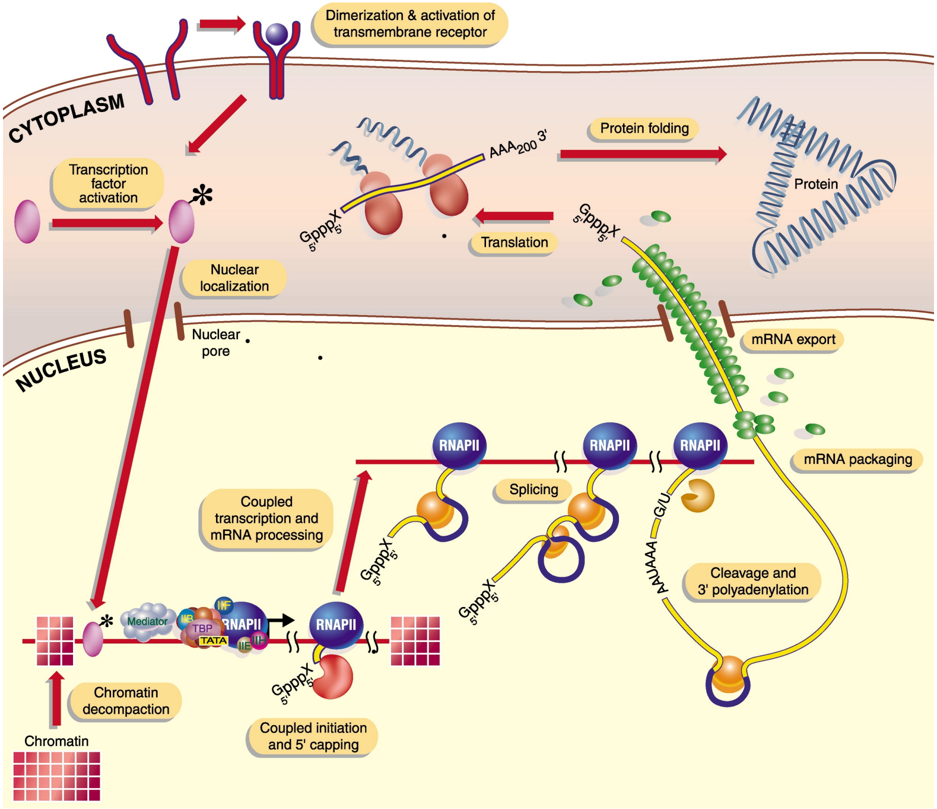

Figure 2.1: Central dogma of molecular biology. The DNA is situated in the nucleus, organized into chromosomes. Every cell must contain the genetic information and the DNA is therefore duplicated before a cell divides (replication). When proteins are needed, the corresponding genes are transcribed into RNA (transcription). The RNA is first processed so that non-coding parts are removed (splicing, alternative splicing) and is then transported out of the nucleus. Outside the nucleus, the proteins are synthesized based upon the code in the RNA (translation). Image reproduced with permission from Woychik and Hampsey (2002).

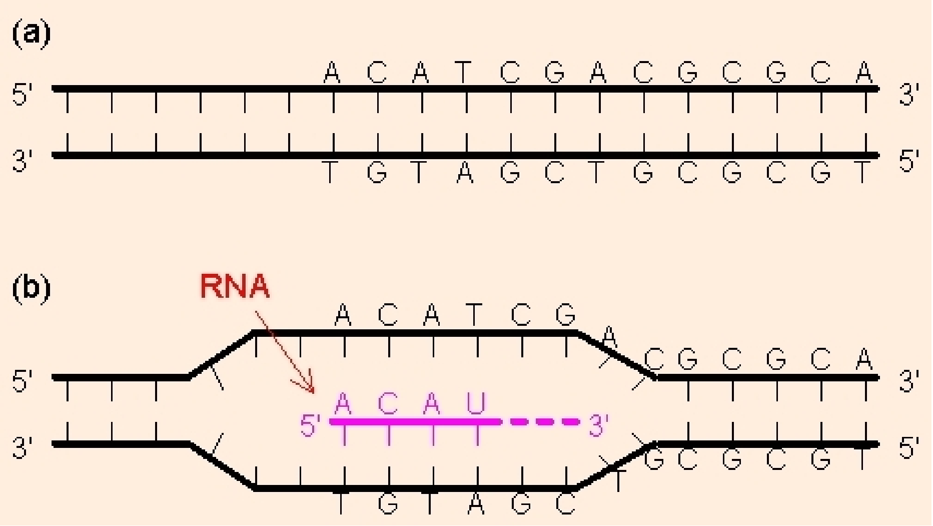

Figure 2.2: Schematic illustration of transcription. (a) DNA before transcription. (b) During transcription, the DNA should unwind so that one of its strand can be used as template to synthesize a complementary RNA. Image reproduced from http://www.web-books.com/MoBio/Free/Ch4B.htm.

2. *Binding of polymerase to the initiation site:* The binding of polymerase to the initiation site is a highly regulated process. It involves several proteins (transcription factors) that bind to the DNA in proximity to the transcription start site called *promoter region.* Some of these proteins bind selectively to regulatory motifs. The motifs in the promoter region are different for different genes. Therefore, a combinatorial binding of the transcription factors to the promoter region implicates the regulation of expression of individual genes. Figure 2.3 illustrates the current knowledge about the transcriptional regulation machinery.

3. *Synthesis of RNA based on the sequence of the DNA template strand:* The synthesis of RNA involves the catalytic activity of enzymes called *RNA polymerases.* In prokaryotes, transcription is carried out by a single type of RNA polymerase or *core* enzyme. In this case, the promoter specificity of RNA polymerase can be altered by different types of sigma factors, which bind to the core enzyme to form a *holoenzyme.* In eukaryotes, most protein-coding genes are transcribed by RNA polymerase II (Pol II).

4. *Termination of synthesis:* The poly-adenylation site marks the end of the transcript. The RNA polymerase changes its elongation capacity as it passes the poly-adenylation signal, which finally leads to the termination of transcription.

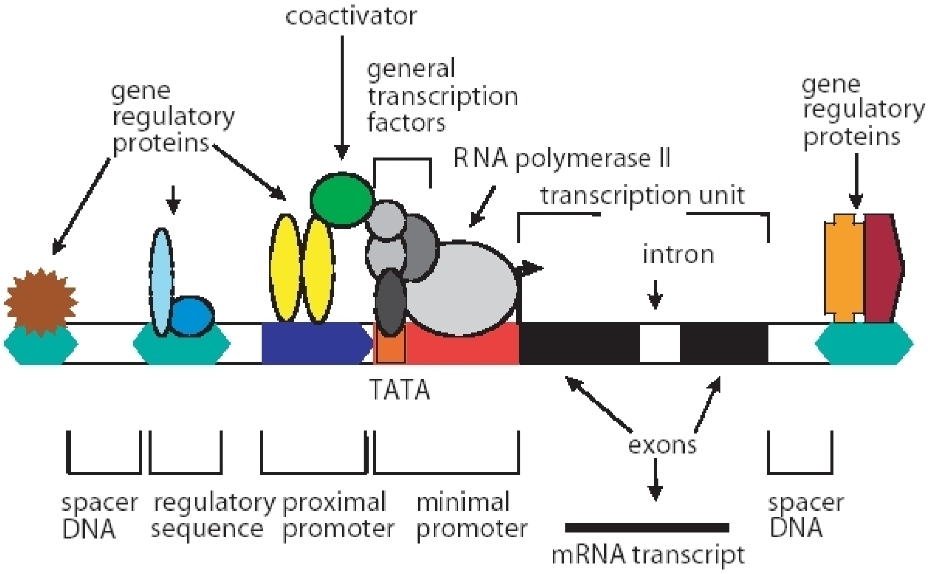

Figure 2.3: Regulation of transcriptional initiation. The minimal promoter is the DNA sequence at which the general transcription factors and RNA polymerase II assemble. The regulatory DNA sequence motifs serve as binding sites for regulatory proteins. These sequences can be located adjacent to the minimal promoter (proximal promoter, unidirectional), far upstream of it (bi-directional, up to several kilobases), or even downstream of the gene. Enhancers are cis-acting sequences that serve as specific binding site for transcription factors to activate transcription. Silencers are cis-acting sequences that serve as specific binding site for transcription factors to inhibit transcription. DNA looping is thought to allow gene regulatory proteins to bind at any of these positions to interact with the proteins that assemble at the promoter. The combination of regulatory proteins and their binding sites are different for each gene. Image reproduced from Villard (2004).

2.1.1 Splicing of pre-mRNA

In most eukaryotic genes, the product of transcription is only a precursor molecule called pre-mRNA. This pre-mRNA is subject to the process of splicing which removes distinct parts of the sequence called *introns*. The introns are marked by *splice signals* that allow their identification by the splicing machinery. Subsequently, the remaining sequence blocks called *exons* are joined together to form the *mature messenger RNA*.

Splicing Signals In all eukaryotes, introns contain and are bordered by highly conserved sequences called the splice signals. The most conserved splice signals bordering the introns are the di-nucleotides GU and AG at the 5' and the 3' end, respectively. These are considered as the consensus splice signals (GU-AG rule). A second splice signal is the presence of an adenosine moiety within the intron (called the branch site). Usually, there is also a U-rich sequence (poly-pyrimidine tract) between the branch site and the 3' splice site. Figure 2.4 summarizes the current knowledge regarding mammalian splice signals. The statistical occurrence rates of the splice signals are comprehensively evaluated in Thanaraj and Clark (2001).

Pre-mRNA splicing is a precisely regulated process. The process begins with the ordered assembly of several small nuclear ribonucleoproteins (snRNP) molecules as well as some non-snRNPs on the pre-mRNA. These proteins identify the splice signals and form the core of *spliceosome* complex. Subsequent assembly of several additional proteins on this core leads to the formation of spliceosome. The entire splicing process involves two major stages:

1. *Formation of the commitment complex:* The process of spliceosome formation starts with the identification of the 5' splice signal, the branch site and the poly-pyrimidine tract by the U1, U2 and U2AF snRNPs respectively. Another snRNP (U5) interacts with the exons surrounding the intron boundaries. This is followed by the binding of other snRNPs (U4 and U6) as well as some non-snRNPs (RNA helicases and SR proteins) to form the core of spliceosome complex. Subsequently, more than 60 additional proteins assemble on this core and form the spliceosome (Stevens et al. (2002), reviewed in Burge et al. (1999), Will and Luhrmann (2001)). Dynamic interactions between the spliceosomal proteins and the pre-mRNA bring the reactive sites in close proximity, thereby creating catalytic sites for trans-esterification reactions. Figure 2.5 illustrates the proteins involved in the formation of the commitment complex.

2. *The trans-esterification reactions:* After the formation of the commitment complex, the cleavage of introns and ligation of exon ends proceeds via a couple of trans-esterification reactions (Figure 2.6). First the 5' exon is cleaved and the 5' end of the intron joins the branch point, creating an intron lariat structure. In

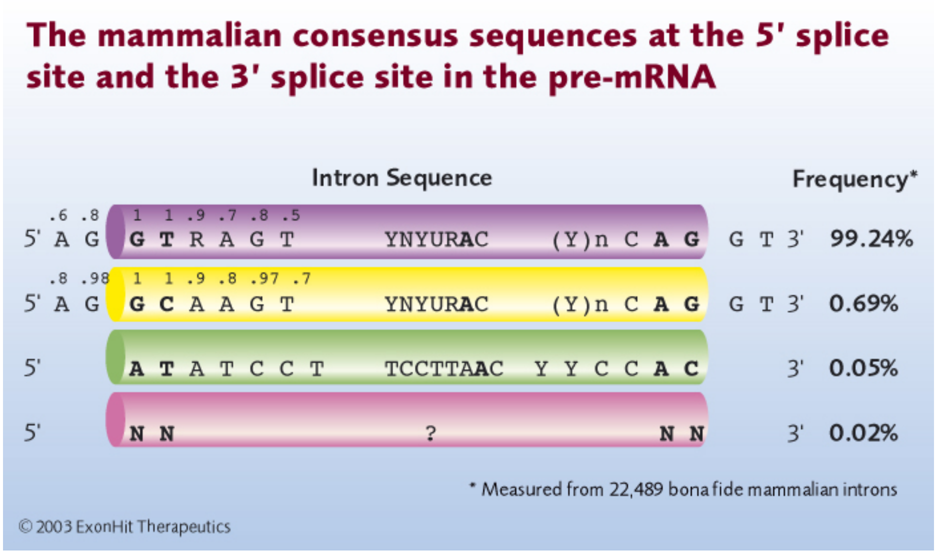

Figure 2.4: The mammalian consensus sequences at the 5' splice site and the 3' splice site in the pre-mRNA. The initial steps in the spliceosomal assembly are directed by several consensus sequences in the pre-mRNA. These sequences are located at the intron-exon junctions and in the intron. The 5' splice site is defined by the consensus sequence - MAG/GURAGU (M = A or C; R = A or G and the / indicates the exon - intron junction). The 3' splice site is defined by three sequence elements going 5' to 3': the branch site (YNYURAC, where A is indicates the adenosine used to form the lariat intermediate structure during splicing; Y = U or C; N = A or G or U or C) the poly-pyrimidine tract, and the 3' splice site consensus (YAG/G; Y = U or C). The branch-point consensus sequence is usually located 18 to 38 nucleotides upstream of the 3' splice site. Image reproduced with permission from http://www.exonhit.com/alternativesplicing/pages/rna_processing/2/index.html.

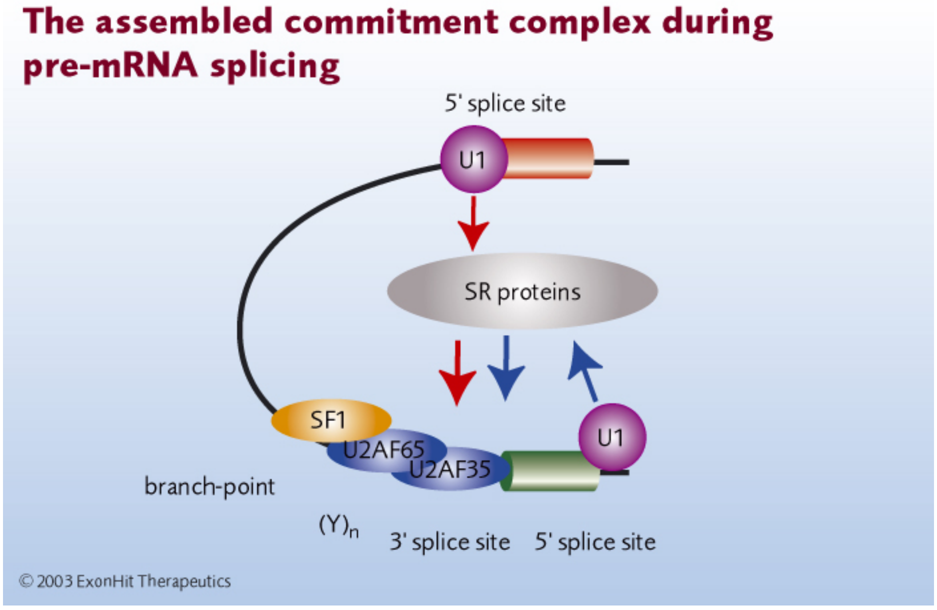

Figure 2.5: The assembled commitment complex during pre-mRNA splicing. This complex can be converted into the active spliceosome and involves the recognition of the 5' splice site by U1 snRNP and the branch-point sequence and 3' splice site by SF1 and U2AF, respectively with the aid of SR proteins. The arrows indicate that interactions may have to occur across introns (intron bridging - red arrows) or across exons (exon bridging - blue arrows) in order to achieve correct pairing of 5' and 3' splice sites. Image reproduced with permission from http://www.exonhit.com/alternativesplicing/pages/rna_processing/2/index.html.

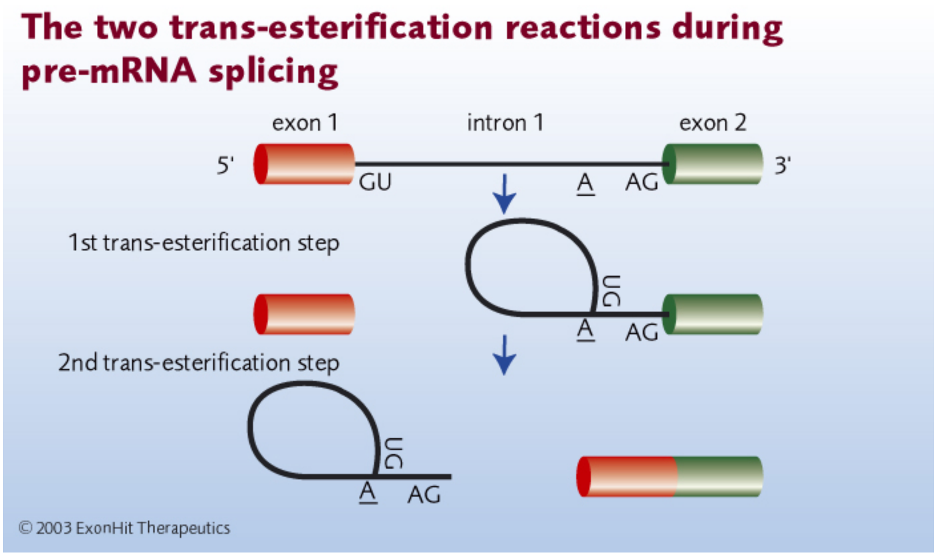
The two trans-esterification reactions during pre-mRNA splicing
exon 1
intron 1
exon 2
5'
3'
GU
A
AG
1st trans-esterification step
UG
A
AG
2nd trans-esterification step
UG
A
AG
© 2003 ExonHit Therapeutics

Figure 2.6: The two trans-esterification reactions during pre-mRNA splicing. The first step involves cleavage at the 5' splice site to yield a 5' exon intermediate (exon 1) with a free 3' OH group. Simultaneously, the 5' end of the intron is joined, by a phosphodiester bond, to the branch-point adenosine residue within the intron. This forms the lariat intermediate containing the intron with the attached 3' exon (exon 2). During the second step, the lariat intermediate is cleaved at the 3' splice site and the two exons are ligated together by a 3' - 5' phosphodiester bond. This results in the two products: spliced exons and the lariat intron. Image reproduced with permission from http://www.exonhit.com/alternativesplicing/pages/rna_processing/2/index.html.

a second step, the free 3' end of the 5' exon connects to the downstream exon leading to exon ligation and subsequent release of the intron sequence.

2.1.2 Alternative splicing

Alternative splicing (Sambrook (1977)) is the surrogate usage of more than one 3' splice site with a particular 5' splice site and vice-versa. The *cryptic* splice sites lead to multiple mature mRNAs from a single pre-mRNA transcript.
Based on the resulting pattern of exon-intron boundaries, the alternative splice events can be classified into four types (Figure 2.7).

1. *Skipped Exons:* This type of alternative splicing includes those events in which one of the transcripts contained an exon that is skipped in another transcript. This class contains *cassette exons* and *mutually exclusive exons* as described recently (Roberts and Smith (2002)).

2. *Multiple Skipped Exons:* A multiple skipped exon event refers to an event in which several consecutive exons are skipped.

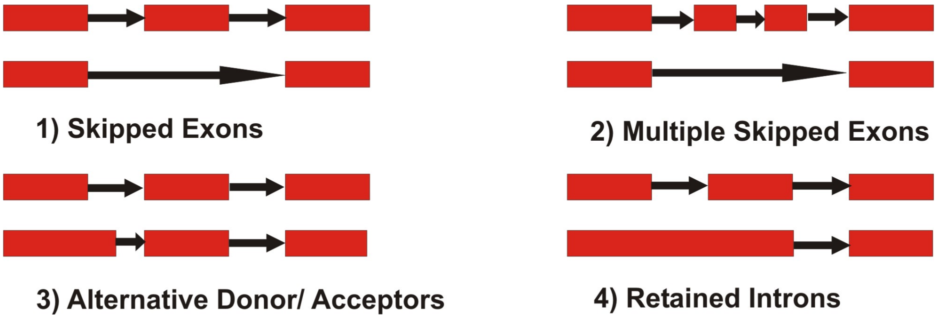

Figure 2.7: Classification of alternative splicing. 1) Skipped exon event in which one of the transcripts has an additional exon. 2) An intron in a transcript is the part of an exon in another transcript. 3) Alternative donor or acceptor sites 4) An intron in a transcript is the part of an exon in another transcript.

3. *Alternative Donor/Acceptor sites:* This category includes instances of alternative splicing in which either the 5' or the 3' splice site is different between the related exons of different transcripts. These events include *competing 5'/3' sites* and *multiple promoter/polyA* (Roberts and Smith (2002)).

4. *Retained Introns:* An event in which an exon in one of the transcripts connects two adjacent exons in another transcript is called a retained intron event.

For a long time, this process of alternative splicing was thought to be an exception. However, over the last few years, it has become apparent that it the process is more of a rule being observed in 40-60% of the genes (Johnson et al. (2003), Kan et al. (2002), Mironov et al. (1999)). Sometimes, the number of isoforms generated by a single gene is enormously large. For example, the Drosophila homolog of human Down syndrome cell adhesion molecule (Dscam) potentially generates more than 38000 isoforms, which may contribute to the specificity of neuronal connectivity (Schmucker et al. (2000)).

Alternative splicing is a highly regulated process with differential inclusion or exclusion of (parts of) exons. One well studied regulatory mechanism for alternative splicing is mediated by the members of SR (serine/arginine rich) protein family. As discussed in the Section 2.1.1, these proteins bind to exons and thereby assist the splicing machinery to recognize them. Together with cis- and trans-occurring splicing *enhancer* and *silencer* sequence elements (Cartegni et al. (2002)), the SR proteins regulate alternative splicing in an environment-dependent manner which lead to either tissue-specific and/or developmental stage-specific splicing (Bourgeois et al. (2004), Sanford et al. (2003)). This control over alternative splicing is critical for proper functioning of the cell. Discrepancies in the splicing regulatory machinery cause several fatal diseases including tumors (Corn and El-Deiry (2002), Caceras and Kornblihtt (2002)).

2.2 Translation

Translation is the process by which the information in mature mRNA is used to synthesize a poly-peptide chain (protein). Unlike DNA to RNA transcription which involves similar molecules (ribonucleic acids), RNA to protein conversion involve another type of molecules called amino acids. Such a conversion involves a genetic code wherein every three nucleotides code for one of the 20 amino acids (Appendix A). The genetic code dictates the synthesis of poly-peptide chain (protein) from a mature mRNA template. The entire process consists of three stages:

1. *Initiation:* The site for translation is the *ribosome* which is made up of protein and ribosomal RNA (rRNA). Translation begins as the small subunit of the ribosome slides down the strand of RNA until it finds the sequence AUG. This recognition implicates the larger subunit of the ribosome to bind to the smaller subunit. Subsequently, the tRNA that has an anticodon corresponding to AUG binds to the active site of ribosome.

2. *Elongation:* Once the translation is initiated, the anticodon of another tRNA molecule base pairs with the next three ribonucleotides that follow AUG. The proximity of the two amino acid residues then implicate a peptide bond between each other. After the formation of this first peptide bond, the ribosome proceeds to the next codon on the RNA molecule. This process is repeated until the termination signal is detected.

3. *Termination:* The three codons (viz. UAA, UAG, UGA) for which there are no tRNA molecules with anticodons that can base pair them are the stop codons. These codons are recognized by a protein *release factor*, which results in the release of the newly formed strand of amino acids from the ribosome. Subsequently, the ribosome splits into its subunits, which are reassembled for another round of protein synthesis.