

Section 1

Fundamentals of Molecular and Cellular Biology

COPYRIGHTED MATERIAL

1

The Central Dogma: From DNA to RNA, and to Protein

Takashi Ohtsuki and Masahiko Sisido

Department of Bioscience and Biotechnology, Okayama University, Japan

Within a single cell – the minimum unit of every living organism – many millions of different types of molecule are working to maintain the cell, to promote its replication, or even to cause its suicide. The bioprocesses conducted within the cell are chemical reactions that proceed under the control of a highly organized network of molecular interactions between relevant biomolecules.

Among these biomolecules, three types of biopolymer are crucial, namely nucleic acids, proteins and polysaccharides. Nucleic acids preserve, replicate and transform the genetic information that serves to design a number of different proteins and low-molecular-weight biomolecules. Proteins function at almost all stages of the bioprocesses, from the birth to the death of a cell. Polysaccharides play important roles in communicating molecular network information and in storing chemical energy. Biopolymer concentrations are regulated to optimum levels for each stage of the bioprocess, but decompose when their roles are complete. This chapter will focus on the molecules and bioprocesses that are related to protein biosynthesis, where DNA is the source of genetic information and the amino acids are the raw materials.

1.1 Chemistry of DNA

Deoxyribonucleic acid (DNA) is a biopolymer that is located inside the nucleus of mammalian cells or in the cytosol of bacterial cells. DNA stores the genetic information that will be converted into the amino acid sequences of protein molecules in the cell.

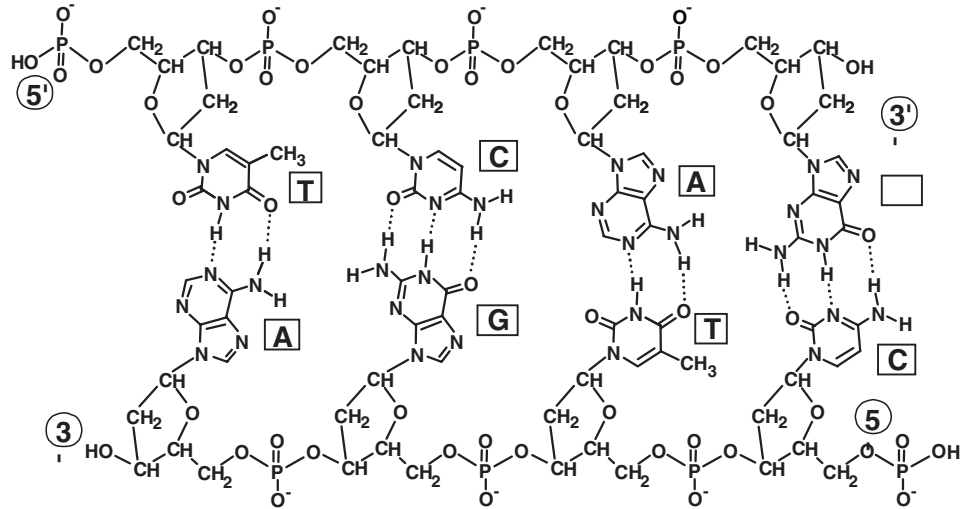


Figure 1.3 Chemical structure of double-stranded DNA

the form of the sequences of nucleobases. The double-helical structure, however, is not absolutely stable, and unfolds at high temperatures or by the action of an enzyme called a helicase.

1.2 Replication of DNA

In order for genetic information to be transferred to the next generation, DNA must first be copied to replicate itself. DNA replication is conducted with an aid of an enzyme called DNA polymerase. The basic chemistry of the replication proceeding inside the enzyme is illustrated in Figure 1.5.

First, the double-helical chain is unfolded and one of the DNA chains is copied to create its complementary chain. The monomer units involved in this polymerization are activated nucleotide units, dATP, dTTP, dGTP and dCTP. The triphosphate unit of the dNTP units is very susceptible to the attack of the 3'-OH group, and forms a diphosphate linkage. Guided

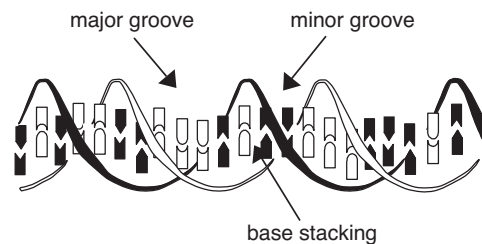


Figure 1.4 Double-helical structure of a complementary pair of two DNA strands

6 Automation in Proteomics and Genomics

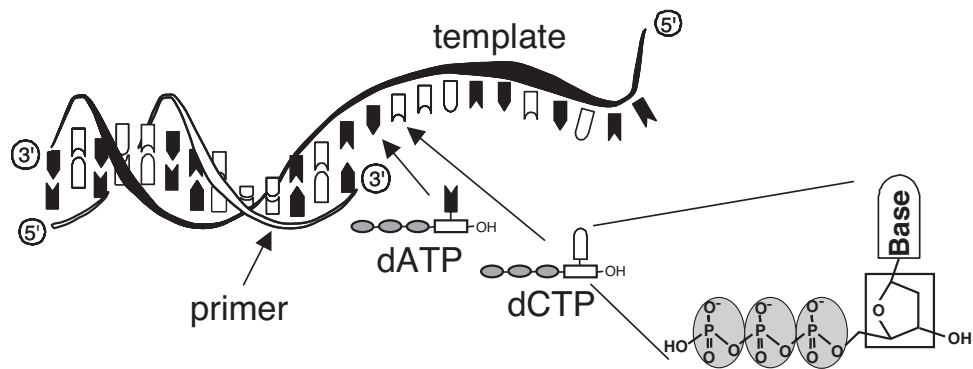


Figure 1.5 Replication of one of the DNA chains to create a complementary chain

by the enzyme, a correct monomer binds to the template DNA chain and reacts with the 3'-OH group of the growing chain. In this way, the new chain grows from the 5' end to the 3' end.

1.3 Transcription from DNA to RNA

Although the stable and inflexible DNA double-helical structure is suitable for the storage of genetic information, its large size necessitates that a smaller, more flexible biopolymer, is used to translate the stored genetic code into proteins. To that end, the base sequences are copied into another type of biopolymer nucleic acid, specifically ribonucleic acid (RNA).

RNA is structurally different from DNA in two ways (see the left part of Figure 1.6). The first difference is that an OH group is attached to the 2'C atom of deoxyribose unit; the 2'-OH derivative is called a ribose unit. The second difference is that a methyl group is removed from the thymine unit to make a uracil unit, U. The introduction of a 2'-OH

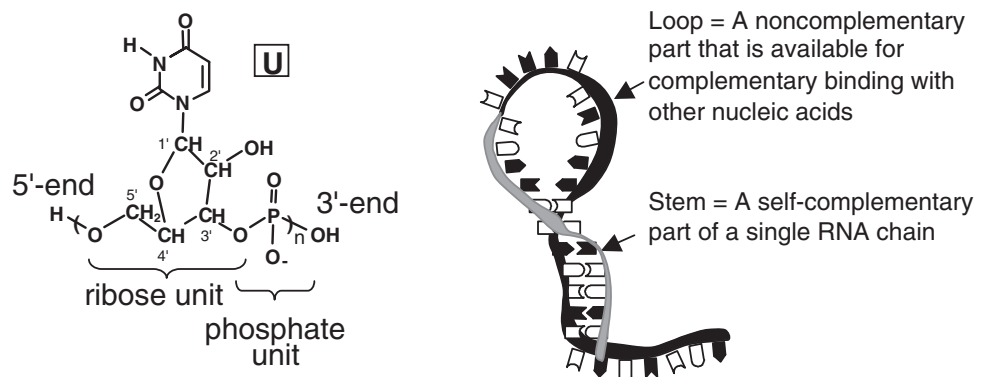


Figure 1.6 Chemical structure of RNA (left) and typical hydrogen-bonded structure of a single RNA chain

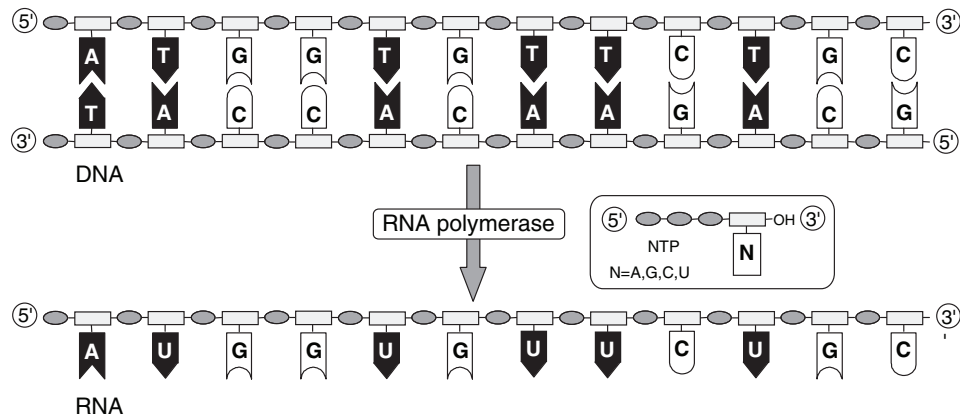


Figure 1.7 Transcription of one of the double-stranded DNA chains to a complementary RNA chain with the aid of RNA polymerase

group causes a small conformational change on the ribose unit such that the RNA chain will favor single-stranded conformations. The single-stranded RNAs, however, often assume an intramolecularly hydrogen-bonded structure, such as a stem-loop structure (see Figure 1.6, right).

Similar to DNA replication, one of the double-stranded DNA chains is copied to a single RNA chain of the complementary nucleobase sequence, except for the alteration of T to U, as shown schematically in Figure 1.7. This procedure is known the transcription process, and is conducted with an enzyme called RNA polymerase. The chemistry of transcription is similar to the replication process, and the monomers are ATP, UTP, GTP and CTP.

1.4 Translation of the Nucleobase Sequence of mRNA to the Amino Acid Sequence of Protein

The information stored in the form of a nucleobase sequence along an RNA chain is translated to an amino acid sequence of a protein, as shown schematically in Figure 1.8. RNAs that serve the translation process are called messenger RNAs (mRNAs). In the translation process, three consecutive nucleobases on a mRNA are taken together and converted to a specific amino acid. The set of three nucleobases is called a codon. As four possibilities (A, U, G and C) exist for each nucleobase, there are $4^3 = 64$ different codons.

Adapter molecules bridge the codons and the amino acids. A class of small RNAs, called transfer RNAs (tRNAs), serve as those adapters. The base sequence of a yeast tRNA that bridges between a codon UUC and an amino acid, phenylalanine, is shown in Figure 1.9.

tRNAs commonly have stem-loop structures with three loops and four stems. Among the loops, the anticodon loop contains three consecutive nucleobases that bond specifically to its complementary codon; thus, a tRNA of a specific anticodon binds to a specific codon on an mRNA. If a particular amino acid is linked to a specific tRNA of specific anticodon, the amino acid will be called up by the codon. In this way, the sequence of nucleobases is translated to the sequence of amino acids.

8 Automation in Proteomics and Genomics

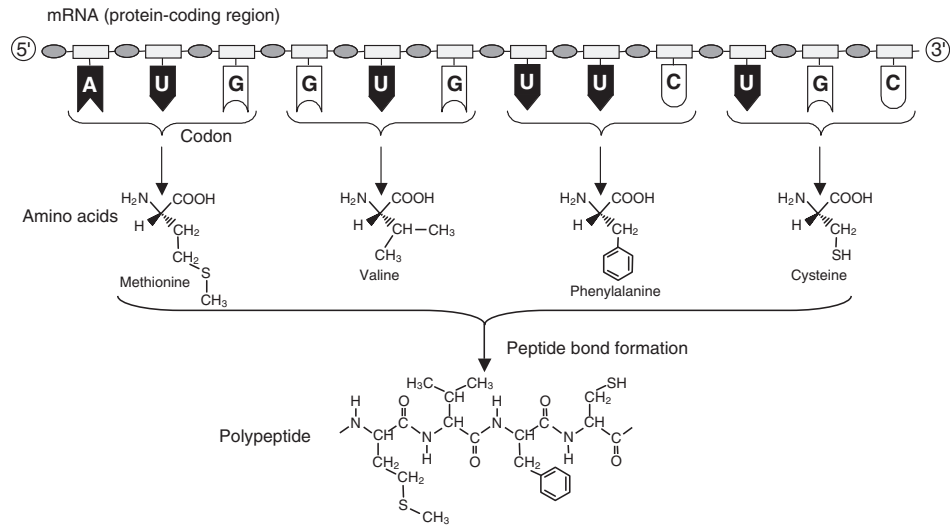


Figure 1.8 Translation of a nucleobase sequence on an mRNA to an amino acid sequence of a protein. A set of three consecutive nucleobases (codon) corresponds to a specific amino acid. The amino acids will be linked together to produce a polypeptide chain

1.5 The Codon Table

A list that correlates between the base sequences of codons and the amino acids is called a codon table (see Figure 1.10). The codon table is common to almost all organisms on the

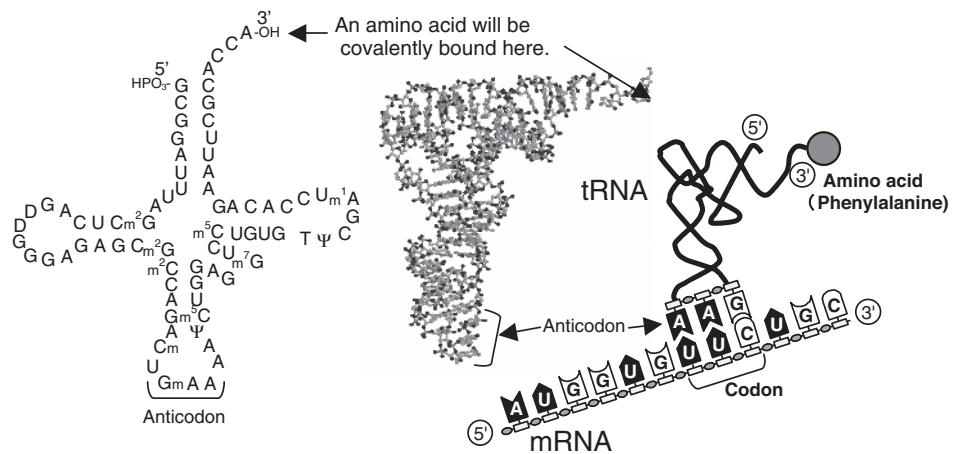


Figure 1.9 Yeast transfer RNA (tRNA) that bridges between a codon UUC and an amino acid, phenylalanine. Nucleobase sequence (left), crystal structure (center) and a schematic illustration of the codon/anticodon pairing. The tRNA contains modified nucleobases, such as m²G, m²C, Cm, Gm, Ψ and D

1st letter	U	C	A	G	2nd letter
U	UUU Phe	UCU	UAU Tyr	UGU Cys	
	UUC	UCC Ser	UAC	UGC	
	UUA Leu	UCA	UAA Stop	UGA Stop	
	UUG	UCG	UAG Stop	UGG Trp	
C	CUU	CCU	CAU His	CGU	
	CUC Leu	CCC Pro	CAC	CGC Arg	
	CUA	CCA	CAA Gln	CGA	
	CUG	CCG	CAG	CGG	
A	AUU	ACU	AAU Asn	AGU Ser	
	AUC Ile	ACC Thr	AAC	AGC	
	AUA	ACA	AAA Lys	AGA Arg	
	AUG Met	ACG	AAG	AGG	
G	GUU	GCU	GAU Asp	GGU	
	GUC Val	GCC Ala	GAC	GGC Gly	
	GUA	GCA	GAA Glu	GGA	
	GUG	GCG	GAG	GGG	

Figure 1.10 The codon table

earth, except for several violations found in mitochondria. The codon table is, therefore, the second basic rule of living organisms.

Because 64 codons correspond to 20 amino acids, there is redundancy in the use of codons. For example, phenylalanine is coded by both UUU and UUC, while leucine is coded by six codons. It must be noted that UUA, UAG and UGA do not correspond to any amino acid, and so are called stop codons. Thus, if they appear on an mRNA, the protein synthesis will cease. Compared with the stop signal, the mechanism of the start of protein synthesis is a little complicated, and is different in bacteria and eukaryotic cells. In prokaryotic bacteria, mRNA has a special region, called the Shine–Dalgarno (SD) sequence, and the first AUG codon after the SD sequence works as the start codon. In eukaryotic cells, the first AUG codon from the 5' terminal of an mRNA is the start codon. In any case, the N-terminal amino acid of a newly synthesized protein will be methionine.

1.6 The Twenty Amino Acids

The types of amino acid that constitute proteins, again, are common to all organisms; the chemical structures of the 20 amino acids are listed in Figure 1.11.

Amino acids are classified into five types, depending on chemical and physical properties of their side groups. The first group (Gly, Ala, Val, Leu, Ile, Met and Pro) has hydrophobic

10 Automation in Proteomics and Genomics

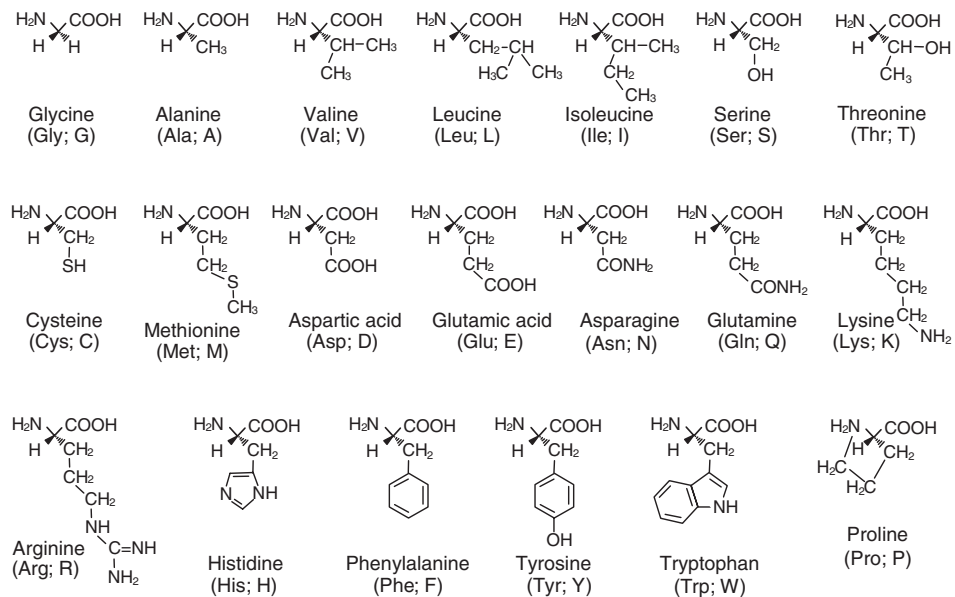


Figure 1.11 Amino acids that constitute proteins

side groups, while the second group (Ser, Cys, Thr, Asn and Gln) has nonionic polar side groups. The third group (Asp and Glu) has anionic side groups, and the fourth (Arg and Lys) has cationic groups. The fifth group (His, Phe, Tyr and Trp) has aromatic groups. Protein conformations depend on the combination of the different groups along a polypeptide chain, such as water-soluble globular proteins, membrane-penetrating hydrophobic proteins, and so on.

1.7 Aminoacylation of tRNA

The process of linking a specific amino acid to a specific tRNA is called the aminoacylation of tRNA, and is governed by a single enzyme, aminoacyl tRNA synthetase (ARS) for each amino acid. For example, phenylalanine (Phe) is charged onto a tRNA that has an anticodon UUU or UUC, with an enzyme PheRS. Aminoacylation consists of two stages, as illustrated in Figure 1.12 (top). First, a particular amino acid is bound to its specific ARS and is activated with adenosine triphosphate (ATP) to form an adenylated amino acid. In the second stage, a particular tRNA is bound to its specific ARS that holds the adenylated amino acid. The latter then reacts with the 3'-terminal OH group of the tRNA to form an ester linkage between the amino acid and the tRNA.

ARS is a 'super' enzyme that recognizes three different substrates: ATP, a specific amino acid, and a specific tRNA. In the first stage of aminoacylation, the formation of an adenylated amino acid activates an amino acid. The mixed anhydride of carboxylic acid and phosphoric acid of the adenylated amino acid is very susceptible to water. Inside the enzyme, however, the mixed anhydride is kept safe, until a correct tRNA is bound in

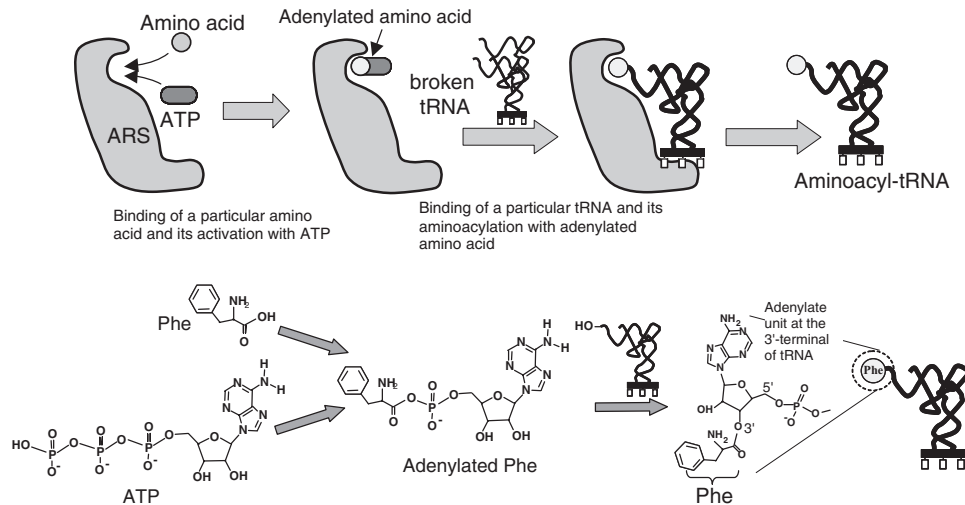


Figure 1.12 Schematic illustration (top) and chemical processes (bottom) of tRNA aminoacylation

proximity so as to induce the aminoacylation. The accuracy of the ARS/amino acid/tRNA selection is very high, and the probability of an erroneous aminoacylation is less than 10^{-4} .

1.8 Protein Synthesis in Ribosomes

Codon/anticodon pairing consists of only three base pairs, and is not strong enough to hold the tRNA/mRNA hybrids. Consequently, aminoacyl tRNAs do not bind to mRNA in solution, even if they have correct anticodons against the codons on mRNA. The codon/anticodon pairing takes place only inside a huge molecular assembly, called a ribosome, which is constructed from RNAs and proteins (Figure 1.13). Inside a ribosome, there are two 'rooms' – one for an aminoacyl tRNA (A site) and the other for a tRNA linked with the growing peptide (P site). There are also two 'tunnels' – one for an mRNA and the other for the growing peptide. Protein synthesis proceeds inside a ribosome, as illustrated in Figures 1.13 and 1.14.

After a tRNA has been aminoacylated with the relevant ARS, it is brought into the A site of the ribosome by the aid of an enzyme, elongation factor-Tu (EF-Tu). In the A site, the aminoacyl tRNA is oriented to locate its amino group in close proximity to the C-terminal ester group of the growing peptide on the tRNA in the P site (see Figure 1.14, top, left). The amino group then attacks the ester group, leading to the formation of a new peptide bond. As the result of this peptide bond formation, the growing peptide transfers to the tRNA in the A site (peptidyl transfer; see Figure 1.14, top, right). The A site tRNA, carrying the growing peptide, is then translocated to the P site, leaving the A site vacant (Figure 1.14, bottom). Finally, the next aminoacyl tRNA will be brought into the vacant A site. This polymerization cycle will be repeated until one of stop codons (UAA, UAG and UGA)

12 Automation in Proteomics and Genomics

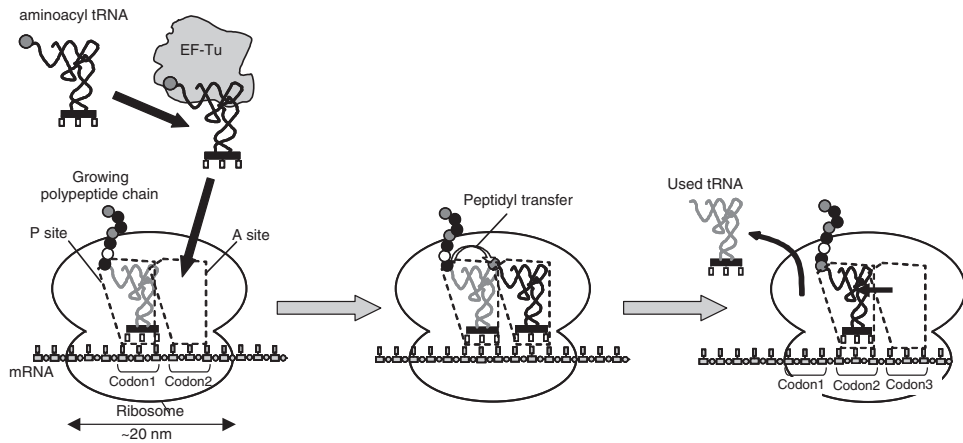


Figure 1.13 Schematic illustration of the course of protein synthesis inside a ribosome

appears on the mRNA. The polypeptide synthesis proceeds at a rate of about two amino acids per second in eukaryotic cells, and about 20 amino acids per second in bacteria.

1.9 The Total Process of Protein Synthesis: 'The Central Dogma'

The entire bioprocess – from DNAs to proteins and from amino acids to polypeptides – is summarized in Figure 1.15. The protein biosynthetic process is essentially the same in all organisms and is referred to as the 'central dogma', although several important differences exist between bacterial and eukaryotic systems.

The central dogma consists of two paths: one path for a flow of information from the nucleobase sequences of DNAs to the amino acid sequences of proteins, and a second path for a flow of materials from amino acids to polypeptides.

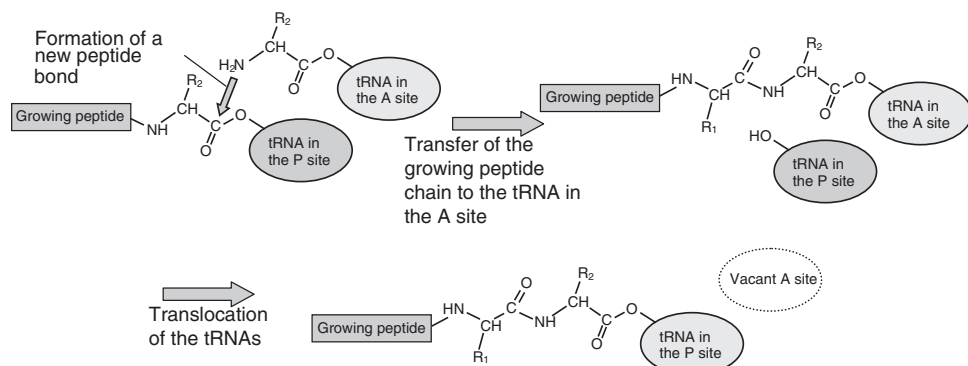


Figure 1.14 Chemistry of protein synthesis occurring inside a ribosome

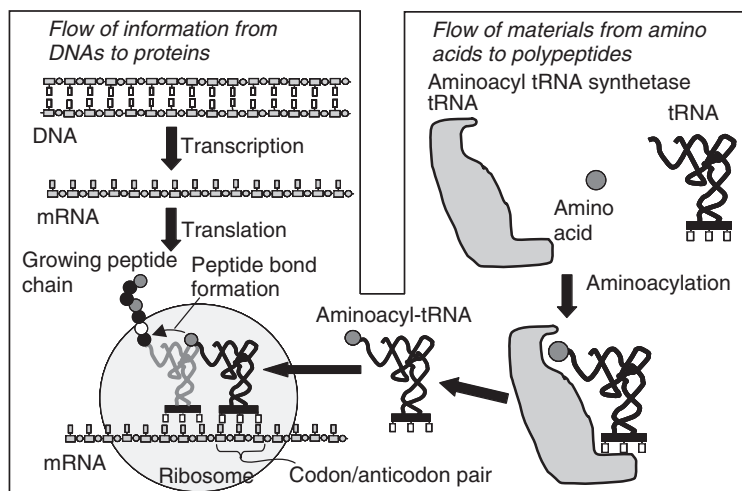


Figure 1.15 The central dogma: The protein biosynthetic process

It is counterintuitive that all organisms from bacteria, to plants and to animals, share essentially the same biosynthetic mechanism, due to the obvious differences in their physical appearances. As no organism lives with only 18 types of amino acid or with six types of nucleobase, it can be deduced that all living organisms are descendants from a single common cell that was comprised of 20 types of amino acid, four types of nucleobase and, essentially, the same protein-biosynthesizing system, as shown in Figure 1.15. Currently, a number of chemists are attempting to expand the central dogma and to create a 'new life' that lives with more than 20 types of amino acid or with more than four types of nucleobase.

1.10 Proteins: Polypeptides with a Variety of Specialty Side Groups that are Spatially Arranged to Achieve Biological Functions

Proteins are constructed from polypeptide chains along which a variety of functional groups are rationally arranged to play individual roles. Unlike most synthetic polymers, the polypeptide main chain is relatively stiff; such rigidity is due to amide groups favoring a planar and trans geometry resulting from the partial shift of an electron from nitrogen to oxygen (Figure 1.16).

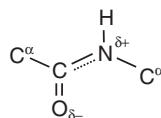


Figure 1.16 Planar and trans geometry of an amide group

14 Automation in Proteomics and Genomics

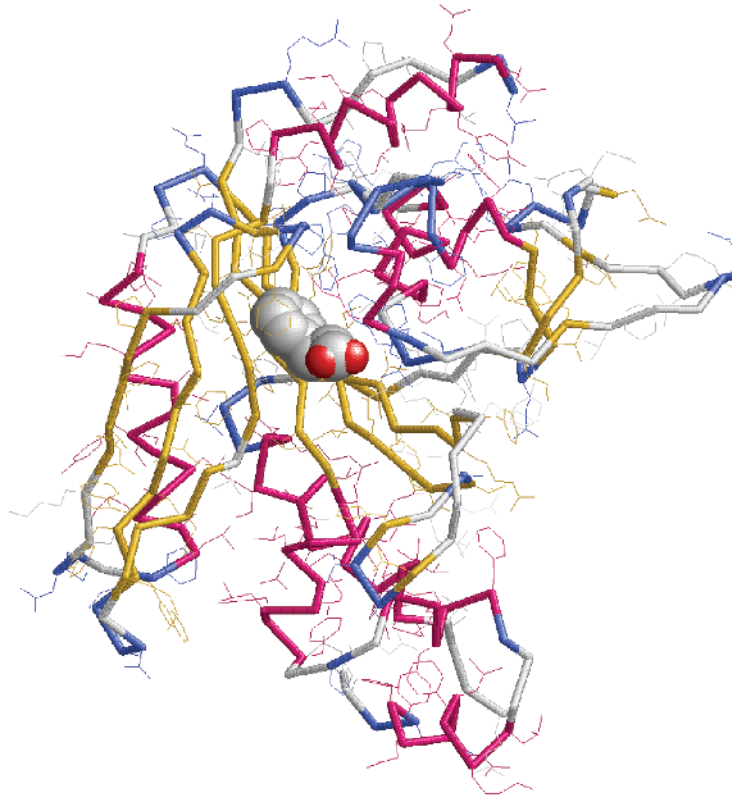


Figure 1.17 Three-dimensional X-ray crystallographic structure of bacterial phenylalanyl tRNA synthetase (PheRS). PheRS consists of two polypeptide chains. In this figure only the phenylalanine binding part (A chain) is shown. The bound phenylalanine is shown by a space-filling model. The main chain is shown by rods. The dark gray portions are in α -helical conformation, while the light gray portions are in β -sheet forms. The side chains are shown with fine lines

Moreover, because of the electronic polarization, the N–H group is an excellent proton donor, while the oxygen atom, in turn, is an excellent proton acceptor. As a result, the amide groups in a polypeptide chain are able to build a strong hydrogen bond network with each other. If the hydrogen bonds were formed between amide groups that are separated by every three α -carbon atoms along a single polypeptide chain, then the latter will take a right-handed α -helical conformation. If hydrogen bonds form to assemble several antiparallel-running chains together, then the polypeptide chains will assume a β -sheet structure.

By combining these structural motifs, such as α -helices and β -sheets, proteins may take a variety of main chain conformations. As an example, a main chain structure of a bacterial phenylalanyl tRNA synthetase (PheRS) is shown in Figure 1.17. The α -helical parts of the main chain are shown in dark gray, and the β -sheet parts in light gray. The bound phenylalanine is presented by a space-filling model.

As amino acid side groups appear in every three atoms along a polypeptide chain, severe crowding is expected between them. Thus, the orientations of the side groups are very constrained and, if they were properly arranged, the side groups would form a three-dimensional space for the effective binding of external molecules, or build a functional region for achieving enzymatic reactions. Figure 1.17 shows how a substrate (phenylalanine, shown by the space-filling model) is bound to its binding site that is made of constrained orientations of the side groups located nearby.

1.11 Genetic Engineering

The central dogma tells us that the amino acid sequences of proteins are determined solely by the nucleobase sequences of protein-coding DNA. Therefore, if new DNA can be synthesized, or if some nucleobases can be substituted with other nucleobases, and the new DNA is introduced into the protein-biosynthesizing system, then new or partially mutated proteins will be created. This technique is known as ‘genetic engineering’, and is widely applied in agricultural, pharmaceutical and medical fields.

In order to introduce new or mutated DNAs into living organisms, for example *Escherichia coli*, a small cyclic double-helical DNA, called a plasmid, is used as the transporter or a vector of the gene (Figure 1.18). The plasmid contains functional units, as depicted in the figure. The new gene is inserted into the protein-coding region by cutting off a portion by restriction enzymes, *EcoRI* and *HindIII* (see Figure 1.18) and pasting a new gene in place of the missing portion by an enzyme called a DNA ligase.

Restriction enzymes cleave double-stranded DNA chains at their specific sites, as typically exemplified for *EcoRI* and *HindIII* in Figure 1.19.

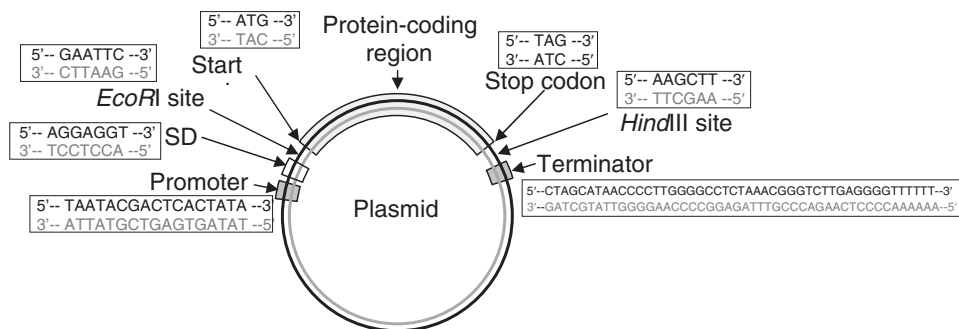


Figure 1.18 Arrangement of functional components along a plasmid. The plasmid is a cyclic double-helical DNA of several thousand base pairs. The promoter sequence determines the start point of RNA polymerization. The Shine–Delgado (SD) sequence determines the point of ribosome attachment. Protein synthesis starts from the start codon (ATG) to one of the stop codons (TAG, TAA, TGA). The terminator sequence determines the end of transcription. The plasmid contains two restriction sites (*EcoRI* site and *HindIII* site in the above example) for inserting new genes. In the above example, the promoter and terminator sequences are taken from those of T7-phage, because of their high efficiencies

16 Automation in Proteomics and Genomics

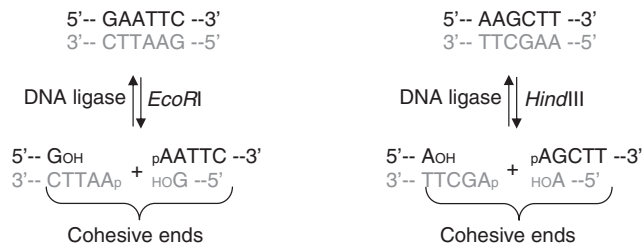


Figure 1.19 DNA cleavage with restriction enzymes (downward arrows) and ligation with DNA ligase (upward arrows). EcoRI cleaves a double-helical DNA at the GAATTC/CTTAAG site. HindIII cleaves at the AAGCTT/TTCGAA site, leaving cohesive ends, respectively. The cohesive ends can be ligated again with the DNA ligase

The cleavage leaves a pair of short complementary chains (cohesive ends) which will be linked again with an enzyme, DNA ligase. Ligation also takes place between the cohesive ends that are produced from different double-stranded DNAs, cleaved by the same type of restriction enzyme. Therefore, if the same set of restriction sites were to exist on a plasmid (Figure 1.20, top) and on a DNA fragment that included the protein-coding region (bottom), the latter would be inserted into the plasmid after cleavage by restriction enzymes, followed by the ligation with DNA ligase (Figure 1.20).

1.12 Large-Scale Production of Engineered Proteins

The complete procedure for obtaining a target protein from the plasmid is illustrated in Figure 1.21. The plasmid inserted with the protein coding region is introduced into

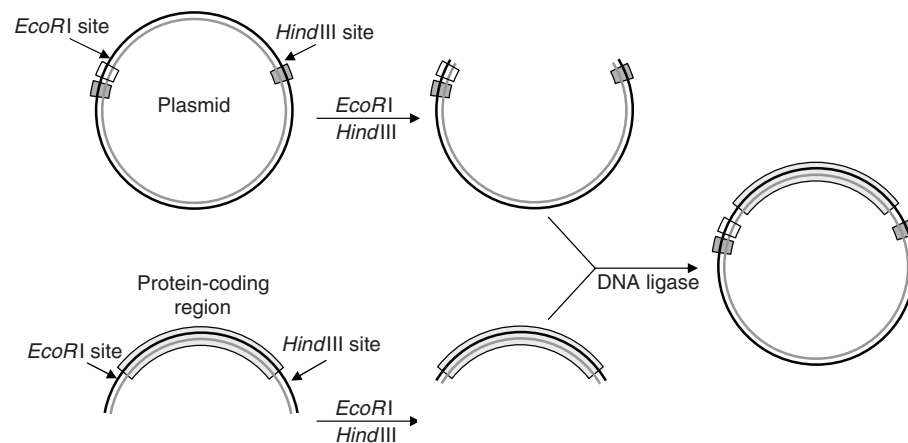


Figure 1.20 Insertion of a protein-coding region on a fragment of double-stranded DNA into a plasmid by the use of a pair of restriction sites on both the plasmid and the DNA

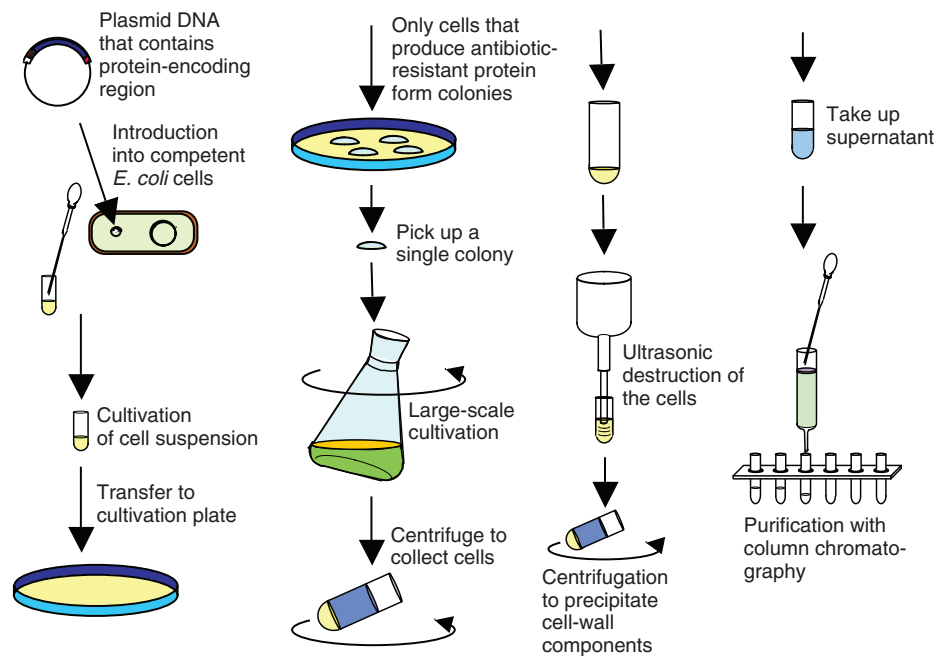


Figure 1.21 Total procedure for the large-scale production of proteins by using *E. coli* transformed with a plasmid

E. coli cells, the cell walls of which are temporarily made permeable to DNAs (competent cells). These transformed cells are cultivated first in suspension, and then transferred onto a cultivation plate with an antibiotic (ampicillin). On the cultivation plate, only those cells that are successfully producing the target protein, together with an ampicillin-resistant protein, can survive and grow to form colonies. Next, one of the colonies is picked up and cultivated in large quantity. After harvesting cells by centrifugation, the cells are lysed by ultrasonic agitation and the insoluble components precipitated by centrifugation. The protein in the supernatant is then purified using column chromatography.

1.13 Cell-Free Protein Synthesis and its Automated Process

Protein synthesis using living cells is advantageous for producing a large quantity of any single type of protein, because the transformed cells can be stored and used repeatedly. However, this approach is not appropriate for synthesizing many different types of protein as, usually, it takes a week (or even longer) to obtain a large quantity of transformed cells. Another drawback of the living cell system is that the expressed proteins often form insoluble aggregates (inclusion bodies) inside the host cells, that are not easily resolved. It is also clear that proteins which are toxic to the host cells cannot be synthesized. Nonetheless, these limitations can be avoided if all of the macromolecules that are functioning in the central dogma are extracted from living cells and then assembled in a test tube to conduct

18 Automation in Proteomics and Genomics

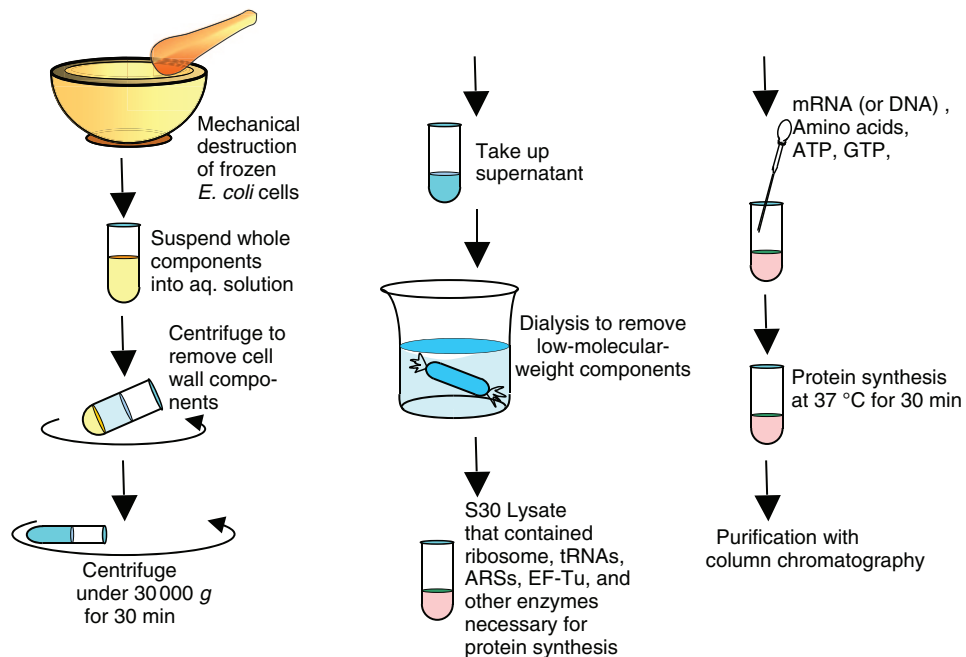


Figure 1.22 Procedure for preparing *E. coli* S30 lysate and cell-free protein synthesis

protein synthesis. The technique is referred to as cell-free protein synthesis, and protein-synthesizing mixtures, taken from *E. coli* for example, are now commercially available.

The procedure for preparing a cell-free protein-synthesizing system (*E. coli* S30 lysate) is illustrated in Figure 1.22. The frozen cells are mechanically destroyed and suspended in aqueous solution. After removal of the insoluble components, the soluble portion is centrifuged at $30\,000 \times g$ for 30 min. The supernatant is then removed and dialyzed against phosphate-buffered saline to remove any low-molecular-weight components. The remaining solution contains tRNAs, ARSs, ribosomes and other enzymes that are necessary for protein synthesis. Following centrifugation, this protein-synthesizing mixture is known as an S30 mixture.

By adding DNA or mRNA and an amino acid mixture, together with energy sources (ATP and GTP) to the S30 mixture, protein synthesis starts rapidly such that within 30 min the target protein is obtained in quantities of approximately $1 \mu\text{g ml}^{-1}$ lysate.

As the cell-free synthesis will cease when one of amino acids or NTPs is exhausted, the materials must be fed continuously in order to continue the synthesis. In addition, waste materials such as diphosphates, nucleotide diphosphates (NDPs) and nucleotide monophosphates (NMPs) must be removed from the reaction mixture. This can be accomplished by using a reaction chamber equipped with an aut feeder separated with a semipermeable membrane, as illustrated in Figure 1.23.

By using such a continuous reaction system the protein yield can be increased to 10-fold that obtained when using a batch system.

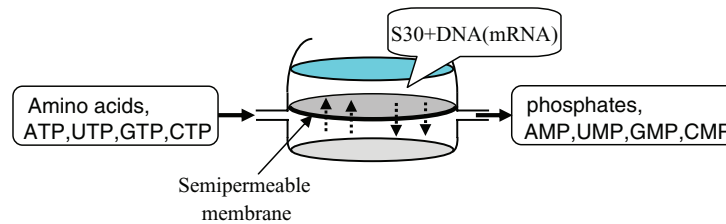


Figure 1.23 Schematic representation of a continuous protein-synthesizing chamber equipped with an autofeeder for amino acids and NTPs

One of the limiting factors of widespread cell-free synthesis is the cost of S30 or other cell lysates. In order to prepare the S30 lysate, a large quantity of *E. coli* cells is required, together with processes that are not suited to large-scale production. However, when the cell-free system becomes less cost-prohibitive, it is poised to become a major protein-producing procedure. Due to the flexibility of the system to synthesize a variety of protein types, the cell-free system is more suited for automated processes than for conventional protein synthesis using living cells.

Acknowledgement

This work was supported by the grants from the National 863 bioinformatics projects under the contract No. 2007AA02Z333, and the Chinese National Science Foundation under the contract No. 20773085, 30870476 and 30770502, as well as the Virtual Laboratory for Computational Chemistry of CNIC, and the Supercomputing Center of CNIC, Chinese Academy of Sciences.

Reference

1. Alberts, B., Johnson, A., Lewis, J. *et al.* (2008) *Molecular Biology of the Cell*, 5th edn, Chapters 5 and 6, Garland Publishing, Inc., New York.

