RIBOSOMES

by Alexander S. Spirin

Institute of Protein Research, Russian Academy of Sciences,
and Department of Molecular Biology, Moscow State University

Preface

This book is based on an advanced course of lectures on ribosome structure and protein biosynthesis that I offer at the Moscow State University. These lectures have been part of a general course on molecular biology for almost three decades, and they have undergone considerable evolution as knowledge has been progressing in this field. The progress continues, and readers should be prepared that some facts, statements and ideas included in the book may be incomplete or out-of-date. In any case, this is primarily a textbook, but not a comprehensive review. It provides a *background* of knowledge and current ideas in the field and gives *examples* of observations and their interpretations. I understand that some interpretations and generalizations may be tentative or disputable, but I hope that this will stimulate thinking and discussing better than if I left white spots.

The book has a prototype: it is my monograph *"Ribosome Structure and Protein Biosynthesis"* published by the Benjamin/Cummings Publishing Company, Menlo Park, California, in 1986. Here I have basically kept the former order of presentation of the topics and the subdivision into chapters. The contents of the chapters, however, have been significantly revised and supplemented. The newly written chapters on translational control in Prokaryotes (Chapter 16) and Eukaryotes (Chapter 17) are added. The chapters on morphology of the ribosome (Chapter 5), ribosomal RNA (Chapter 6) and cotranslational folding and transmembrane transport of proteins (Chapter 18) are completely rewritten in the co-authorship with Dr. V. D. Vasiliev, Prof. A. A. Bogdanov and Prof. V. N. Luzikov, respectively. The concluding chapter on general principles of ribosome structure and function is appended.

The literature references in this book, as in the previous one, are given mainly for teaching purposes, so that the reference lists at the end of each chapter are far from complete. To give an insight into the histories of discoveries I cited preferentially pioneer studies in the fields discussed. To provide information on the present state of knowledge, I have referred the reader to some of the recent publications. In addition, many illustrations, specifically those which are borrowed from other authors, are supplied with corresponding references. The book contains also many original illustrations made due to invaluable help of my colleagues at the Institute of Protein Research, Pushchino, especially P. G. Kuzin, A. Kommer, and V. A. Kolb. The assistance of L. N. Rozhanskaya, the secretary, M. G. Dashkevitch and V. V. Sosnovsky, Computers and Communication Department, and T. B. Kuvshinkina and M. S. Shelestova, Scientific Information Department, in preparing the manuscript is also greatly appreciated.

I am grateful to all my colleagues, as well as other scientists, who have read parts of the manuscripts and made their comments.

Alexander S. Spirin

Pushchino and Moscow,
July 1998

PART I

HISTORICAL AND FUNDAMENTAL INTRODUCTION

Chapter 1

PROTEIN BIOSYNTHESIS: SUMMARY AND DEFINITIONS

The proteins of all living cells are synthesized by *ribosomes.* The ribosome is a large macromolecule consisting of ribonucleic acids *(ribosomal RNAs)* and proteins; it has a complex asymmetric quaternary structure. In order to synthesize protein, the ribosome must be supplied with (1) a program determining the sequence of amino acid residues in the polypeptide chain of a protein, (2) the amino acid substrate from which the protein is to be made, and (3) chemical energy. The ribosome itself plays a catalytic role and is responsible for forming peptide bonds, i.e. for the polymerization of amino acid residues into the polypeptide chain.

The program that sets the sequence of amino acid residues in a polypeptide chain comes from *deoxyribonucleic acid* (DNA), i.e. from the cell genome. Sections of the double-stranded DNA, which are called genes, serve as templates for synthesizing single-stranded RNA molecules. The synthesized RNA species are complementary replicas of just one of the DNA chains and therefore are faithful copies of the nucleotide sequence of the other DNA chain. This process of gene copying, accomplished by the enzyme RNA polymerase, is called *transcription.* In eukaryotic cells, and to a lesser extent in prokaryotic cells, nascent RNA may undergo a number of additional changes called *processing;* as a result, certain parts of the nucleotide sequence may be excised from RNA, and in some cases altered (edited). The mature RNA becomes associated with the ribosomes and serves as a program, or template, which determines the amino acid sequence in the synthesized protein. This template RNA is usually called *messenger RNA* (mRNA). In other words, the flow of information from DNA to ribosomes is mediated by gene transcription and RNA processing, resulting in the formation of mRNA.

In the eukaryotic cell the production of mRNA, that is transcription and most events of processing, is compartmentalized in the nucleus. At the same time all functioning ribosomes are localized in the cytoplasm. Hence, the *transport* of mRNA from the nucleus to the cytoplasm is a necessary step in the flow of information from DNA to ribosomes. In Prokaryotes, as well as in eukaryotic cytoplasmic organelles (mitochondria and chloroplasts), DNA and ribosomes are present in the same compartment, so that the ribosomes can reach mRNA and start to synthesize proteins during transcription; this is the so-called *coupled transcription-translation.*

Proteins consist of amino acids. Free amino acids, however, are not used in the synthetic machinery of the ribosome. To become a substrate for protein synthesis, an amino acid must be *activated* by coupling with the adenylic moiety of ATP and then *accepted* by (covalently linked to) a special RNA molecule called *transfer RNA* (tRNA); this process is performed by the enzyme aminoacyl-tRNA synthetase. The resulting aminoacyl-tRNA is used by the ribosome as a substrate for protein synthesis, and the energy of the chemical bond between the amino acid residue and tRNA is used for forming a peptide bond. Thus, the activation of amino acids and formation of aminoacyl-tRNAs provide both material and energy to protein synthesis.

Using mRNA as a program and aminoacyl-tRNAs as energy-rich substrates, the ribosome *translates* genetic information from the nucleotide language of mRNA into the amino acid language of polypeptide chains. In molecular terms this implies that while moving along the mRNA, the ribosome consecutively selects appropriate aminoacyl-tRNA species from the medium. The specificity of the aminoacyl residue of a corresponding aminoacyl-tRNA selected by the ribosome is defined by the combination of nucleotides in a corresponding stretch of mRNA associated with the ribosome. This brings us to the problem of *genetic coding,* i.e. the question of nucleotide combinations that determine, or code, each of the 20 natural amino acids. These combinations are known to be nucleotide triplets, which are called *codons.*

Hence, the movement of the ribosome along the mRNA chain (or, in other words, the passing of mRNA through the ribosome) establishes a temporal order of entering the various aminoacyl-tRNA species into the ribosome. This order depends on the sequence of coding nucleotide combinations (codons) along the mRNA. The aminoacyl residue of each selected aminoacyl-tRNA is being attached covalently to

a growing polypeptide chain by the ribosomal machinery. Deacylated tRNA is released by the ribosome into solution. In each act of aminoacyl-tRNA selection and deacylated tRNA release an additional energy, in the form of GTP hydrolysis, is consumed by the ribosome. All this results in the step-by-step formation of the polypeptide chain, according to the program of mRNA.

A short historical review article, citing 311 references from 1897 to 1980, was presented by P. Siekevitz & P. C. Zamecnik in 1981 ("Ribosomes and protein synthesis", *J. Cell Biol.* 91: 53s-65s).

The general model of protein biosynthesis outlined above is schematically presented in Fig. 1.1.

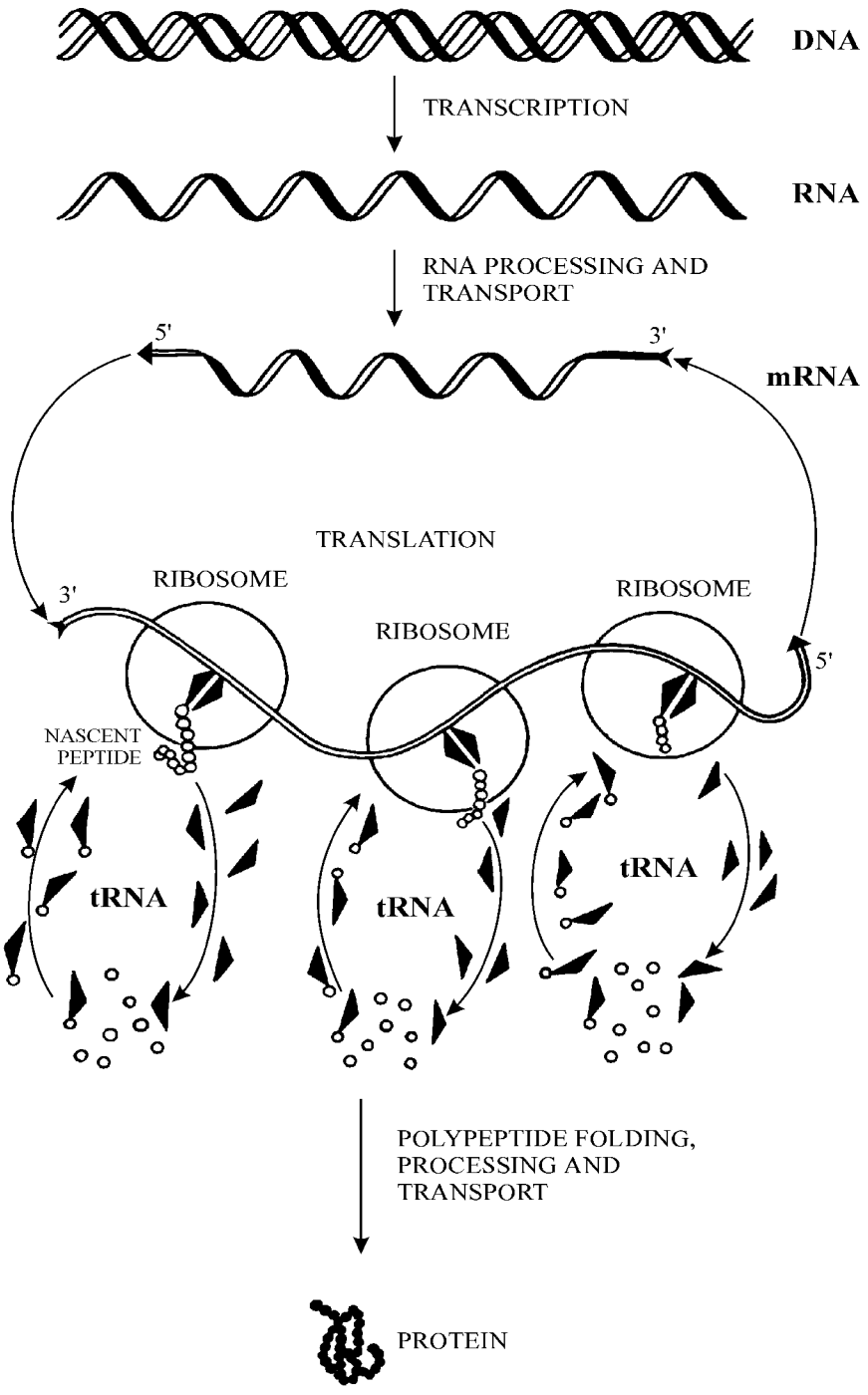

Figure 1.1. General schematic model of protein biosynthesis (DNA → RNA → protein).

Chapter 2

MESSENGER RNA AND THE GENETIC CODE

2.1. Discovery of mRNA

After the discovery and final recognition of the genetic function of DNA (Avery, MacLeod & McCarty, 1944; Hershey & Chase, 1952; Watson & Crick, 1953), it rapidly became clear that DNA itself does not serve as a direct template for protein synthesis. In addition, a number of early observations suggested that ribonucleic acid is closely connected to cellular protein synthesis (Caspersson *et al.*, 1941; Brachet, 1941–1942). These ideas were developed and resulted in the concept that RNA is the intermediate responsible for the transfer of genetic information from DNA to proteins; in particular it has been suggested that RNA serves as a template upon which amino acid residues are polymerized (DNA → RNA → protein) (see Crick, 1959).

This conceptual advance coincided with the discovery of protein-synthesizing ribonucleoprotein particles of the cell which were later called ribosomes (see Chapter 4). It had also been established that the RNA of these particles accounted for the main bulk of cellular RNA. Hence, it was naturally assumed that genes are transcribed into ribosomal RNA species, which in turn serve as templates for protein synthesis. This led to a "one gene - one ribosome - one protein" hypothesis.

During 1956 to 1958, in order to test this hypothesis, a comparative analysis of DNA and RNA base composition in a large number of microorganisms was conducted (Belozersky & Spirin, 1958). DNA base compositions can be rather different in different groups of microorganisms, and it was hypothesized that if the above formulation of a "DNA→RNA→protein" model was correct, the base composition of total RNA would strongly correlate with the DNA base composition in bacteria. The experimental results, however, were unexpected. Despite great differences of DNA base composition in various bacterial species, the composition of total RNA was found to be similar in all of the studied bacteria, and did not mimic DNA base composition. These results implied that the bulk of cellular RNA, i.e. most likely ribosomal RNA, could not serve as a direct informational intermediate between DNA and proteins.

At the same time, RNA base composition was shown to vary slightly for different bacterial species, and to be positively correlated with the base composition of DNA. The conclusion based on this correlation was that cells may contain a special *minor RNA fraction* which imitates DNA base composition and could possibly serve as an intermediate between genes and protein-synthesizing particles (Belozersky & Spirin, 1958).

Earlier Volkin and Astrachan (1956) studied RNA synthesis in bacteria infected with DNA-containing T2 bacteriophage. Bacterial protein synthesis ceases soon after infection, and the entire cellular protein-synthesizing machinery is switched over to producing phage proteins. Most of the cellular RNA does not undergo any change during this process, but the cell begins to synthesize a small fraction of metabolically unstable short-lived RNA, the nucleotide composition of which is similar to the base composition of phage DNA.

Several years later, in 1961, the minor RNA fraction, termed *DNA-like RNA,* was separated from the total cellular RNA. Its function as messenger, carrying information from the DNA to the ribosomes, was demonstrated in the direct experiments of Brenner, Jacob, and Meselson (1961), and those of Gros, Watson, and co-workers (1961); similar observations have been made by Spiegelman and associates (1961). It has been demonstrated that DNA-like RNA formed after the T4 phage infection binds to the preexisting host ribosomes (no new ribosomes are synthesized after phage infection), and the ribosomes associated with the phage-specific RNA synthesize the phage proteins. This RNA could be detached easily from the ribosomes *in vitro* without destroying the particles. It has been shown that this RNA is indeed complementary to one of the phage DNA chains.

On the basis of their results on genetic regulation in bacteria, Jacob and Monod (1961) advanced the idea that a special short-lived RNA transfers information from genes to ribosomes and serves as a direct template for protein synthesis. The term *messenger RNA* was accepted in all subsequent studies.

2.2. Deciphering the Code

The first step after the discovery of mRNA (1956–1961) was to elucidate the code by which amino acid sequences of proteins are written in the nucleotide sequences of mRNA and correspondingly in the nucleotide sequence of one of the two DNA chains (see Gamov, Rich & Ycas, 1956). Even before the discovery of mRNA, theoretical considerations led to the assumption that each amino acid had to be coded by a combination of at least three nucleotides. Indeed, proteins are composed of 20 sorts of natural amino acids (Fig. 2.1), whereas nucleic acids contain only 4 types of nucleotide residues; the nitrogenous bases of nucleic acids are adenine (A), guanine (G), cytosine (C), and either uracil (U) for RNA or thymine (T) for DNA. It was obvious that one nucleotide could not code for one amino acid (4 *vs.* 20). There could be 16 dinucleotide combinations, or doublets, a number again insufficient to code for 20 amino acids. Thus, the minimal number of nucleotide residues in a combination coding for one amino acid had to be three; in other words, amino acids most probably had to be coded by the *nucleotide triplets.* The number of possible triplets is 64, more than enough for the coding of 20 amino acids.

There were two possible explanations for excessive triplets: either only 20 triplets are "meaningful", i.e. may code for one or another amino acid, while the other 44 are nonsense ones, or amino acids may be coded by more than one triplet, in which case the code would be *degenerate.*

Furthermore, the triplet code could be overlapping when a given nucleotide is part of three strongly overlapping or two less overlapping coding triplets; alternatively, it could be nonoverlapping when independent coding triplets are adjacent to each other in the template nucleic acid or are even separated by noncoding nucleotides. The observation that point mutations (i.e. changes of a single nucleotide in the nucleic acid molecule) usually lead to a change of only one amino acid in the corresponding protein provided evidence against the idea of an overlapping code. Moreover, the overlapping code would inevitably result in the possible neighbors of a given amino acid residue being restricted, a situation that has never been observed in actual protein sequences. Therefore a *nonoverlapping* cod appeared more likely.

Finally, it had to be demonstrated whether the coding triplets were separated by noncoding residues, or commas, or whether they were read along the chain without any punctuation; in other words, whether the code was *comma-free* or not. The comma-free case leads to the problem of the reading frame of the template nucleic acid: only a strict triplet-by-triplet readout from a fixed point on the polynucleotide chain could result in an unambiguous amino acid sequence.

The classic experiments of Crick, Brenner and associates published at the end of 1961 established that the code is triplet, degenerate, nonoverlapping, and comma-free. In these experiments, numerous mutants were obtained in the rII region of the T4 bacteriophage gene B using chemical agents which produced either insertions or deletions of one nucleotide residue during DNA replication. Proflavine and other acridine dyes were used for this purpose. Nucleotide insertions or deletions close to the gene origin resulted in a loss of gene expression. By recombining different mutant phages in *Escherichia coli* cells, phenotypic revertants showing normal gene expression were obtained. An analysis of the revertants demonstrated that gene expression was restored if the region with the deletion was located near the region with the insertion, or *vice versa.* Gene expression could also be restored if two additional insertions (or deletions) were introduced near the region with the initial insertion (or, respectively, deletion). The following conclusions were drawn: (1) Insertion or deletion of a single nucleotide at the beginning of the coding region appeared to result in a loss of all the coding potential of the corresponding gene instead of simply a point mutation; the inactivation could be the result of a shift of the reading frame. (2) Deletion or insertion located close to the initial insertion or deletion, respectively, restored the coding potential of the sequence because the original reading frame was restored. (3) Three, but no fewer, closely located insertions or deletions also restored the initial coding potential of the nucleotide sequence. From the results of these experiments, it follows that the code is triplet, and that triplets are read sequentially without commas from a strictly fixed point in the same frame. These experiments also provided additional evidence that the code is degenerate: if many of the 64 possible triplets were nonsense ones, it was highly probable that at least one nonsense triplet appeared in the region between the insertion and deletion or between the three insertions where the readout occurs with a shift of frame; this would lead to an interruption of the polypeptide chain synthesis.

Deciphering the nucleotide triplets also began in 1961 when Nirenberg and Matthaei discovered the coding properties of synthetic polyribonucleotides in cell-free translation systems. The possibility of

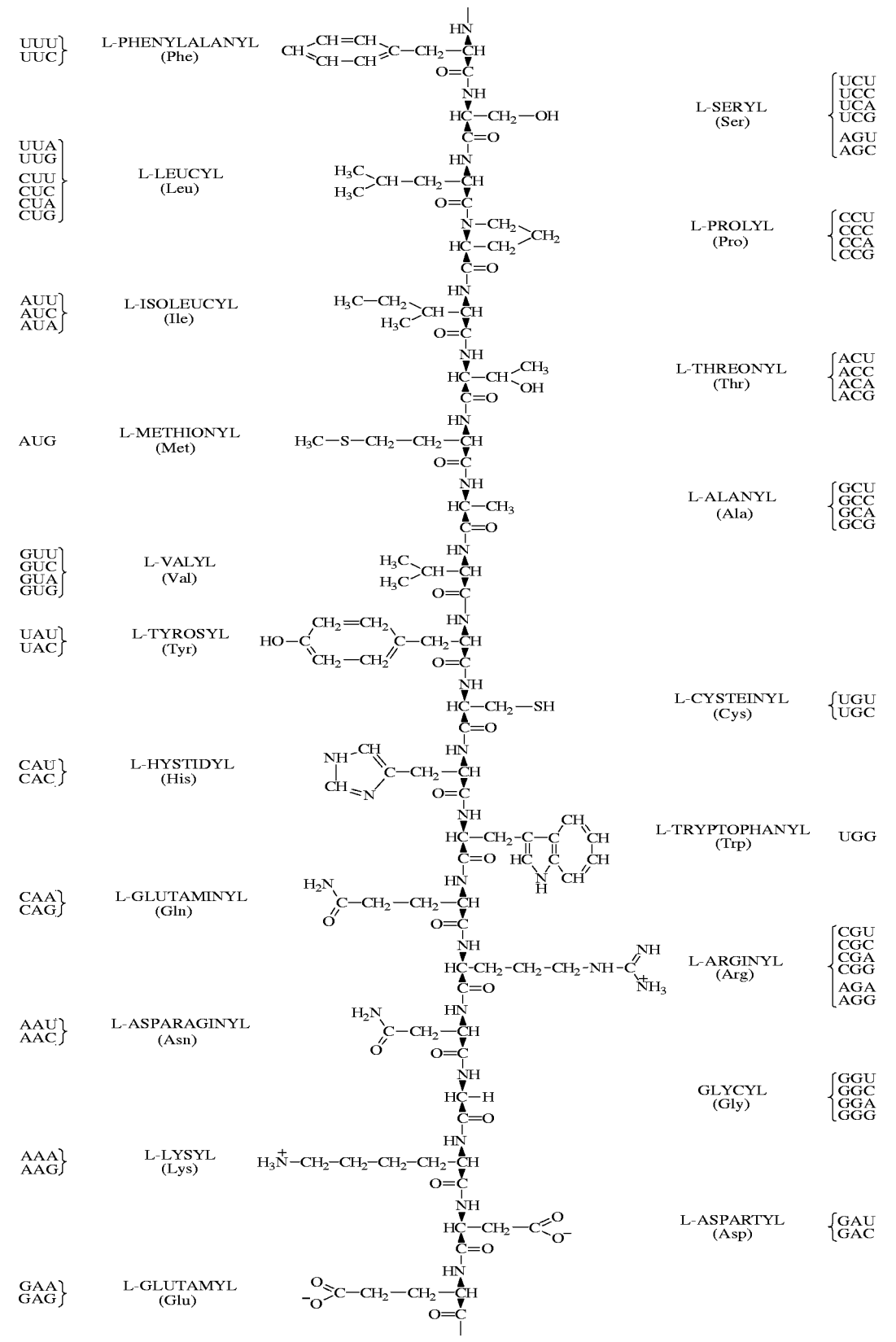

Figure 2.1. Amino acid residues from which proteins are synthesized, and the corresponding codons.

preparing synthetic polyribonucleotides of various compositions using a special enzyme, polynucleotide phosphorylase, was first demonstrated by Grunberg-Manago and Ochoa several years earlier (1955). The composition of polynucleotides synthesized in the system that they described depended only on the selection of ribonucleoside diphosphates supplied as substrates; homopolynucleotides such as polyuridylic acid, polyadenylic acid, and polycytidylic acid prepared from UDP, ADP, and CDP, respectively, were the simplest polyribonucleotides synthesized. Using poly(U) as a template polynucleotide for *E. coli* ribosomes, Nirenberg and Matthaei (1961) demonstrated that this template directs synthesis of polyphenylalanine. It has been concluded that the triplet UUU codes for phenylalanine. Similarly, experiments with polyadenylic and polycytidylic acids have shown that AAA codes for lysine, and CCC for proline.

Further elucidation of the genetic code was based on the use of synthetic statistical heteropolynucleotides of a different composition, which was set by the number and ratio of substrate nucleoside diphosphates in the polynucleotide phosphorylase reaction (Nirenberg *et al.*, 1963; Speyer *et al.*, 1963). Thus, it was demonstrated that the statistical poly(U, C) copolymer directed the incorporation of four amino acids into the polypeptide chain; these were phenylalanine, leucine, serine, and proline. If the U-to-C ratio in the polynucleotide was 1:1, then all four amino acids were incorporated into the polypeptide with equal probabilities. If the U-to-C ratio was 5:1, the probabilities of amino acid incorporation were as follows: Phe > Leu = Ser > Pro. Thus phenylalanine should be coded by triplets consisting of three U or of two U and one C. Leucine and serine are coded by triplets consisting of two U and one C or of two C and one U. Proline is coded by triplets consisting of three C or of two C and one U. Unfortunately, this approach could provide only the composition of the coding triplets, not their nucleotide sequence, since the nucleotide sequence of the template polynucleotide used was statistical.

Due to the invention of a new technique by Nirenberg and Leder (1964), the nucleotide sequences of the coding triplets were soon determined. They found that individual trinucleotides possessed coding properties: after association with the ribosome they supported the selective binding of aminoacyl-tRNA species with the ribosome. For example, UUU and UUC triplets stimulated the binding of phenylalanyl-tRNA, UCU and UCC the binding of seryl-tRNA, CUU and CUC the binding of leucyl-tRNA, and CCU and CCC the binding of prolyl-tRNA. By 1964, methods for synthesizing trinucleotides with the desired sequence were available. In the subsequent two years a wide variety of trinucleotides were tested and, as a result, virtually the whole code was deciphered (Fig. 2.2).

The end of the story was marked by the use of synthetic polynucleotides with a regular nucleotide sequence as templates in the cell-free ribosomal systems of polypeptide synthesis. Methods allowing regular polynucleotides to be synthesized have been developed by Khorana, who has also verified the genetic code by directly using these polynucleotides as templates (Khorana *et al.*, 1966). In complete agreement with the previously established code dictionary, the use of poly$(UC)_n$ as a template resulted in the synthesis of a polypeptide consisting of alternating serine and leucine residues, while poly$(UG)_n$ directed synthesis of the regular copolymer with alternating valine and cysteine residues. Poly$(AAG)_n$ directed the synthesis of three homopolymers: polylysine, polyarginine, and polyglutamic acid.

2.3. Some Features of the Code Dictionary

The complete code dictionary is given in Fig. 2.2. Of the 64 triplets termed *codons,* 61 are meaningful or sense ones: they code for 20 amino acids of natural polypeptides and proteins. Regularly three codons – UAG ("amber"), UAA ("ochre"), and UGA ("opal") – normally do not code for amino acids and therefore are sometimes called *nonsense codons.* The nonsense triplets play an important part in translation, since in mRNA these codons serve as signals for the termination of polypeptide chain synthesis; at present they are usually referred to as *termination* or *stop codons.*

At the same time UGA triplet may also code for the 21st amino acid of a number of proteins, selenocysteine (Chambers *et al.*, 1986; Zinoni *et al.*, 1987). This, however, requires the presence in mRNA of an additional structural element, either immediately adjacent to UGA from its 3'-side (in the case of Prokaryotes), or located beyond the coding sequence, in the 3'-proximal untranslated region of mRNA (in Eukaryotes) (see Chapter 10, Section 10.2.2).

As seen from Fig. 2.2, the degeneracy of the code does not extend to all 20 main amino acids. Two amino acids, methionine and tryptophan, are coded by one codon each, i.e. by AUG and UGG, respectively. On the contrary, three amino acids, specifically leucine, serine, and arginine, have six codons

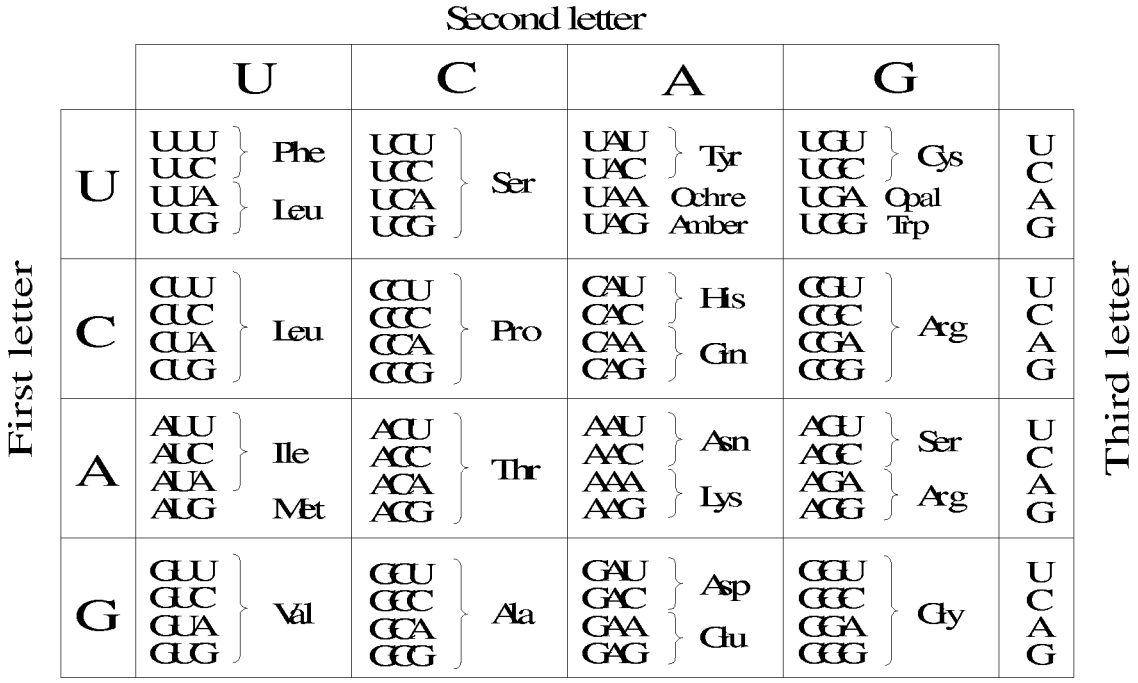

Figure 2.2. Codon dictionary (F.H.C. Crick, *Cold Spring Harbor Symp. Quant. Biol.* **31,** 1–9, 1966).

each. The remaining amino acids, with the exception of isoleucine, are coded either by two or by four codons; only isoleucine is coded by three codons.

It should be emphasized that the triplets coding for a given amino acid differ in most cases only in the third base. Only when the amino acid is coded by more than four codons do differences occur in the first and second positions of the triplet as well. A group of four codons differing only in the third nucleotide and coding for one and the same amino acid is often called the *codon family.* The code dictionary contains eight such codon families: for leucine, valine, serine, proline, threonine, alanine, arginine, and glycine.

The code presented in Fig. 2.2 is universal for the protein-synthesizing systems of most bacteria and for the cytoplasmic extraorganellar protein-synthesizing systems of multi-cellular Eukaryotes, i.e. animals, fungi, and plants.

2.4. Deviations from the Universal Code

By the end of the 1970s and during the 1980s it was discovered that the universality of the genetic code is not absolute, and some exceptions are possible (Barrell *et al.*, 1979; Yamao *et al.*, 1985). Among living organisms, now two genera of eubacteria, *Mycoplasma* and *Spiroplasma,* are known to have two codons for tryptophan, the universal UGG and the "neighboring" UGA, which is a stop codon in other organisms. In one genus of Ciliates (Protozoa), *Euplotes,* UGA codes for cysteine. Two other universal stop codons, UAA and UAG, were reported to code for glutamine in other genera of Ciliates (*Tetrahymena, Paramecium, Stylonicia, Oxytricha*) and in at least one genus of unicellular green algae (*Acetabularia*). Also, in some yeast (*Candida*) the universal leucine codon CUG codes for serine. The known cases of variations in the genetic code are summarized in Table 2.1 (see Watanabe & Osawa, 1995). Further exceptions of the universal genetic code may be discovered in future, especially in unicellular Eukaryotes (Protozoa, algae and fungi).

Organelles of eukaryotic cells, including mitochondria, possess their own protein-synthesizing systems. The protein-synthesizing systems of animal and fungal (but not plant) mitochondria typically show a number of significant deviations from the universal code (Table 2.2). Tryptophan in these

mitochondria is coded by both UGG and UGA; UGA is therefore not used as a termination codon. In mitochondria of all Vertebrates, most (but not all) Invertebrates and some fungi the universal isoleucine codon AUA codes for methionine, so that methionine is determined there by two triplets, the universal AUG and the "neighboring" AUA. The triplets AGA and AGG do not code for arginine in mitochondria of most animals; they are stop codons in vertebrate mitochondria and codons for serine in mitochondria of many Invertebrates (Echinoderms, insects, mollusks, nematodes, Platyhelminthes). In yeast mitochondria (*Saccharomyces, Torulopsis*) the whole codon family CUU, CUC, CUA, and CUG codes for threonine but not for leucine, although in other fungi, such as *Neurospora* and *Aspergillus,* these codons correspond to leucine as given by the universal code.

Table 2.1. Variations in eubacterial and in nuclear genetic code from "universal" genetic code.

Organism	UGA (Stop)	UAA UAG (Stop)	CUG (Leu)
Eubacteria:			
Mycoplasma	Trp	—	—
Spiroplasma	Trp	—	—
Yeasts:			
Candida	—	—	Ser
Ciliates			
Tetrahymena	—	Gln	—
Paramecium	—	Gln	—
Stylonicia	—	Gln	—
Oxytricha	—	Gln	—
Euplotes	Cys	—	—
Unicellular green algae			
Acetabularia	—	Gln	—

Table 2.2. Variations in mitochondrial genetic code.

Organism	UGA Stop	AUA Ile	AAA Lys	AAA AGG Arg	CUN Leu	UAA Stop
Vertebrates	Trp	Met	—	Stop	—	—
Tunicates	Trp	Met	—	Gly	—	—
Echinoderms	Trp	—	Asn	Ser	—	—
Arthropods	Trp	Met	—	Ser	—	—
Molluscs	Trp	Met	—	Ser	—	—
Nematodes	Trp	Met	—	Ser	—	—
Platyhelminths	Trp	—	Asn	Ser	—	Thr?
Coelenterates	Trp	ND	ND	—	ND	ND
Yeasts	Trp	Met	—	—	Thr	—
Euascomycetes	Trp	—	—	—	—	—
Protozoa	Trp	—	—	—	—	—

ND = not determined; — = same as universal code.
After K.Watanabe and S. Osawa, in "tRNA: Structure, Biosynthesis, and Function" (D. Söll and U. RajBhandary, eds.), ASM Press, Washington DC, 1995.

2.5. Structure of mRNA

2.5.1. Primary Structure

In contrast to DNA, messenger RNA, as well as other cellular RNA species, is a *single-stranded* polynucleotide. It consists of four kinds of linearly arranged ribonucleoside residues – adenosine (A), guanosine (G), cytidine (C), and uridine (U) – sequentially connected by phosphodiester bonds between the 3'-position of the ribose of one nucleoside and the 5'-position of the adjacent one (Fig. 2.3). The terminal nucleoside, the 5'-position of which does not participate in forming the internucleotide bond, is referred to as the 5'-end of RNA. The terminal nucleoside with free 3'-hydroxyl is referred to as the 3'-end. It is accepted practice to read and write RNA nucleotide sequences from the 5'- to the 3'-end, i.e. in the direction of the internucleotide phosphodiester bond from the 3'-position to the 5'-position of the neighbor (3'-P-5' bond direction). This direction corresponds to the polarity of mRNA readout by the ribosome.

The terminal 5'-position in natural mRNAs is always substituted. In prokaryotic organisms this end is either simply phosphorylated (Fig. 2.3) or carries the triphosphate group. Eukaryotic mRNAs generally have a special group, the so-called *cap,* at the terminal 5'-position (Furuichi & Miura, 1975; Furuichi *et al.*, 1975). The cap is the N'-methylated residue of guanosine 5'-triphosphate linked with the 5'-terminal nucleoside by the 5'–5' pyrophosphate bond (Fig. 2.4). Eukaryotic cells possess a special system including guanylyl transferase and methyl transferase, enzymes that are responsible for mRNA capping. In addition, the capping is usually accompanied by methylation of the 2'-hydroxyl group of ribose and the base in the 5'-terminal nucleoside adjacent to the cap. Often the 5'-terminal residue in mRNA is a purine nucleoside, either G or A.

The 3'-terminal hydroxyl of natural mRNA remains unsubstituted. Thus, this end possesses two hydroxyl groups in *cis*-position (*cis*-glycol group) (see Fig. 2.3).

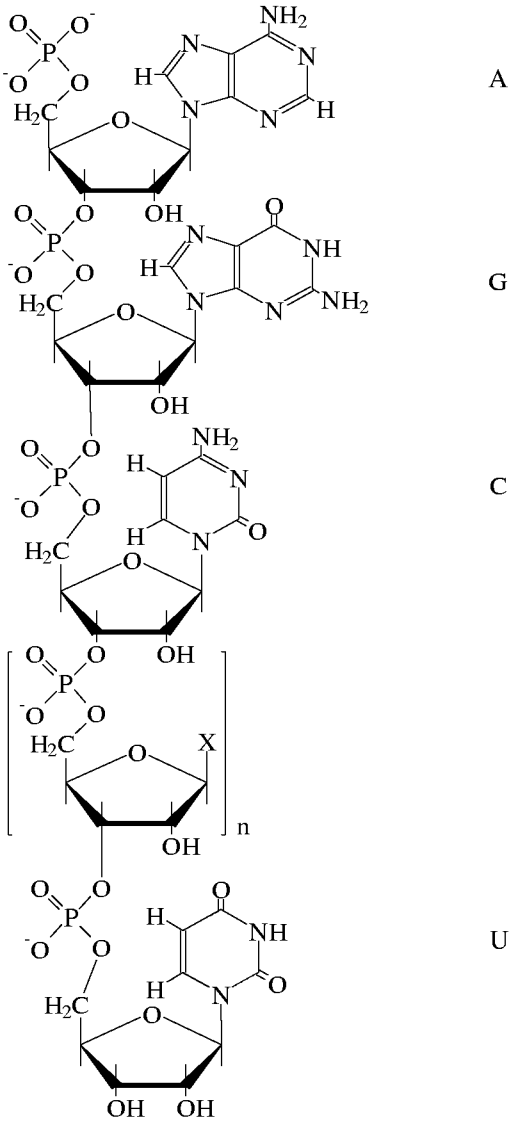

Figure 2.3. Nucleotide residues in RNA.

2.5.2. Functional Regions

The physical length of the mRNA chain is always greater than the length of its coding sequence. The coding sequence includes only part of the total mRNA length. The first codon is preceded by a noncoding (untranslated) 5'-terminal sequence (5'-UTR) the length of which varies for different mRNAs. Furthermore, the terminal codon is never located at the 3'-end of an mRNA chain, but is always followed by a noncoding 3'-terminal sequence (3'-UTR). In addition, most eukaryotic mRNAs contain a long noncoding sequence of adenylic acid residues at their 3'-end. This poly(A) tract (tail) is added to mRNA after the end of transcription by a special enzyme, polyadenylate polymerase.

Identifying the factors that determine the starting point of the coding nucleotide sequence within an mRNA chain is an important problem. Each polypeptide is known to begin with a N-terminal methionine

residue, and therefore the first codon in the coding sequence should be that of methionine. In most cases AUG, and less frequently GUG or UUG (in Prokaryotes), play the role of the *initiation codon* (see Chapter 15). The codon AUG codes for methionine both when it is the first codon of the mRNA coding sequence and when it occurs in internal positions. The codon GUG, however, codes for valine in internal positions and for the initiator methionine only if it occupies the first position in the coding sequence. The same is true for codon UUG coding for leucine in internal positions. In some exceptional cases, AUU or AUA in Prokaryotes and ACG or UUG in Eukaryotes may also serve as initiation codons for the first methionine in the chain. The identification of the initiation codons, however, does not solve the starting point problem of the coding sequence. The difficulty is that by no means every AUG (the more so GUG or UUG) triplet becomes an initiation codon. Generally, translation cannot be initiated from internal AUG, GUG, or UUG triplets. If an mRNA chain is scanned from its 5'-end, AUG as well as GUG and UUG triplets may be found repeatedly both in frame with the subsequent coding sequence and out of frame, but they cannot initiate translation. Finally, many AUG, GUG, and UUG triplets located within the coding sequence but out of the reading frame fortunately do not initiate synthesis of erroneous polypeptides. Thus, in contrast to all other codons, both sense and nonsense ones, the choice of a given codon as an initiation point depends not only on the codon structure, i.e. its nucleotide composition and sequence, but also on the position of the codon in the mRNA. Certain structural elements in mRNA confer the capacity to serve as initiation codon to a given AUG (or GUG, or UUG). Specifically the nucleotide sequence preceding the initiation codon, as well as the particular secondary and tertiary structures of this mRNA region, are vital for the corresponding triplet to be exposed as an initiation codon (Chapter 15).

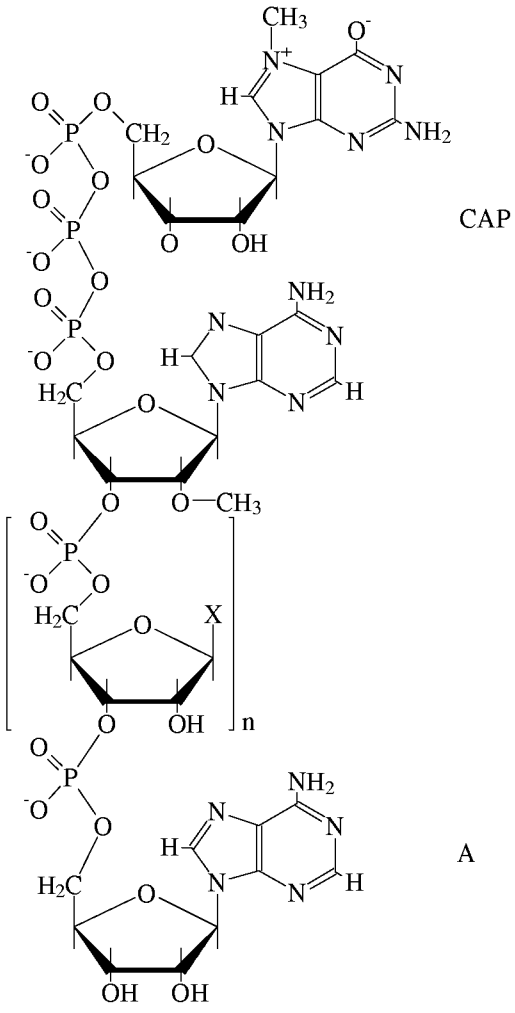

Figure 2.4. Cap structure at the 5'-end of eukaryotic mRNA.

A given mRNA polynucleotide chain does not necessarily contain just one coding sequence. In prokaryotic mRNAs it is common for one polynucleotide chain to contain coding sequences for several proteins. Such mRNAs are usually called *polycistronic mRNAs.* (This term comes from the word *cistron,* which S. Benzer introduced as an equivalent of a gene). Different coding sequences *(cistrons)* within a given mRNA chain are usually separated by internal noncoding sequences. Such an internal noncoding sequence begins from the termination codon of the preceding cistron. The next cistron begins from an initiation codon such as AUG (or GUG).

In contrast to Prokaryotes, in eukaryotic organisms mRNAs are as a rule *monocistronic,* i.e. they code for just one polypeptide chain. The eukaryotic mRNA coding sequence is flanked both at the 5'-end and at the 3'-end by noncoding (untranslated) sequences (5'- and 3'-UTRs), the 3'-UTR being typically very long (comparable with the length of the coding sequence). It has already been mentioned that the vast majority of eukaryotic mRNAs have also poly(A) tracts of various length at the 3'-end. The 5'-end is usually modified by the cap (Fig. 2.4), which appears to be essential for the association between the mRNA and the ribosome prior to initiation.

It is appropriate to emphasize here that the mechanisms responsible for searching for the initiation codon in prokaryotic and eukaryotic translation systems are different. Prokaryotic ribosomes form a complex with mRNA and recognize the initiation codon independently of the 5'-end; it is for this reason that they can initiate from internal sites in the