The Genetic Code

Proteins are the end products of most information pathways. A typical cell requires thousands of different proteins at any given moment. These must be synthesized in response to the cell's current needs, transported (targeted) to their appropriate cellular locations, and degraded when no longer needed. Many of the fundamental components and mechanisms used by the protein biosynthetic machinery are remarkably well conserved in all life-forms from bacteria to higher eukaryotes, indicating that they were present in the last universal common ancestor (LUCA) of all extant organisms.

An understanding of protein synthesis, the most complex biosynthetic process, has been one of the greatest challenges in biochemistry. Eukaryotic protein synthesis requires more than 70 different ribosomal proteins; 20 or more enzymes to activate the amino acid precursors; a dozen or more auxiliary enzymes and other protein factors for the initiation, elongation, and termination of polypeptides; perhaps 100 additional enzymes for the final processing of different proteins; and 40 or more kinds of transfer and ribosomal RNAs. Overall, almost 300 different macromolecules cooperate to synthesize polypeptides. Many of these macromolecules are among the most

abundant to be found in any cell. Some are organized into the complex three-dimensional structure of the ribosome.

To appreciate the central importance of protein synthesis, consider the cellular resources devoted to this process. Protein synthesis can account for up to 90% of the chemical energy used by a cell for all biosynthetic reactions. Every bacterial, archaeal, and eukaryotic cell contains from several to thousands of copies of many different proteins and RNAs. The 15,000 ribosomes, 100,000 molecules of protein synthesis–related protein factors and enzymes, and 200,000 tRNA molecules in a typical bacterial cell can account for more than 35% of the cell's dry weight.

Despite the great complexity of protein synthesis, proteins are made at exceedingly high rates. A polypeptide of 100 residues is synthesized in an *Escherichia coli* cell (at 37 °C) in about 5 seconds. Synthesis of the thousands of different proteins in a cell is tightly regulated, so that just enough copies are made to match the current metabolic circumstances. To maintain the appropriate mix and concentration of proteins, the targeting and degradative processes must keep pace with synthesis. Research is gradually uncovering the finely coordinated cellular choreography that guides each protein to its proper cellular location and selectively degrades it when it is no longer required.

The study of protein synthesis offers another important reward: a look at a world of RNA catalysts that may have existed before the dawn of life "as we know it." Elucidation of the three-dimensional structures of ribosomes, beginning in 2000, has given us an increasingly detailed look at the mechanics of protein synthesis. It has also confirmed a hypothesis first put forward by Harry Noller two decades earlier: proteins are synthesized by a gigantic RNA enzyme.

27.1 The Genetic Code

Three major advances set the stage for our present knowledge of protein biosynthesis. First, in the early 1950s, Paul Zamecnik and his colleagues designed a set of experiments to investigate where in the cell proteins are synthesized. They injected radioactive amino acids into rats and, at different time intervals after the injection, removed the liver, homogenized it, fractionated the homogenate by centrifugation, and examined the subcellular fractions for the presence of radioactive protein. When hours or days were allowed to elapse after injection of the labeled amino acids, *all* the subcellular fractions contained labeled proteins. However, when only minutes had elapsed, labeled protein appeared only in a fraction containing small ribonucleoprotein particles. These particles, visible in animal tissues by electron microscopy, were therefore identified as the site of protein synthesis from amino acids, and later were named ribosomes **(Fig. 27-1)**.

Paul Zamecnik, 1912–2009
[Source: Archives and Special Collections, Massachusetts General Hospital.]

The second key advance was made by Mahlon Hoagland and Zamecnik when they found that amino acids were "activated" for protein synthesis when incubated with ATP and the cytosolic fraction of liver cells. The amino acids became attached to a heat-stable soluble RNA of the type that had been discovered and characterized by Robert Holley, and later called transfer RNA (tRNA), to form **aminoacyl-tRNAs**. The enzymes that catalyze this process are the **aminoacyl-tRNA synthetases**.

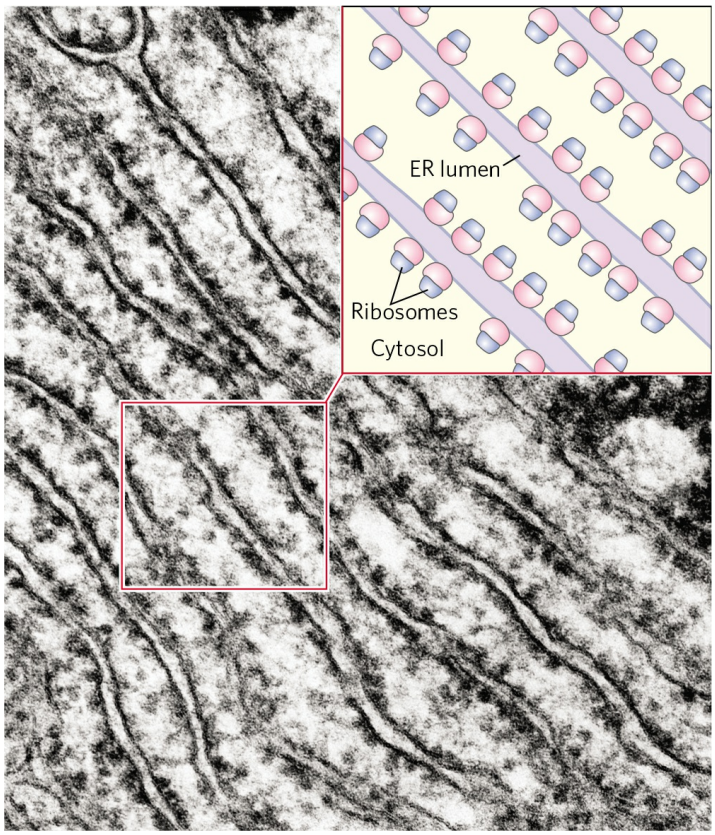


FIGURE 27-1 Ribosomes and endoplasmic reticulum. Electron micrograph and schematic drawing of a portion of a pancreatic cell, showing ribosomes attached to the outer (cytosolic) face of the endoplasmic reticulum (ER). The ribosomes are the numerous small dots bordering the parallel layers of membranes.
[Source: Joseph F. Gennaro Jr./Science Source.]

The third advance resulted from Francis Crick's reasoning on how the genetic information encoded in the 4-letter language of nucleic acids could be translated into the 20-letter language of proteins. A small nucleic acid (perhaps an RNA) could serve the role of an adaptor, with one part of the adaptor molecule binding a specific amino acid and another part recognizing the nucleotide sequence encoding that amino acid in an mRNA **(Fig. 27-2)**. This idea was soon verified. The tRNA adaptor, the same molecule that activates the amino acid for peptide bond formation, also "translates" the nucleotide sequence of an mRNA into the amino acid sequence of a polypeptide. The overall process of mRNA-guided protein synthesis is often referred to simply as **translation**.

These three developments soon led to recognition of the major stages of protein synthesis and ultimately to elucidation of the genetic code that

specifies each amino acid.

The Genetic Code Was Cracked Using Artificial mRNA Templates

By the 1960s, it was apparent that at least three nucleotide residues of DNA are necessary to encode each amino acid. The four code letters of DNA (A, T, G, and C) in groups of two can yield only $4^2 = 16$ different combinations, insufficient to encode 20 amino acids. Groups of three, however, yield $4^3 = 64$ different combinations.

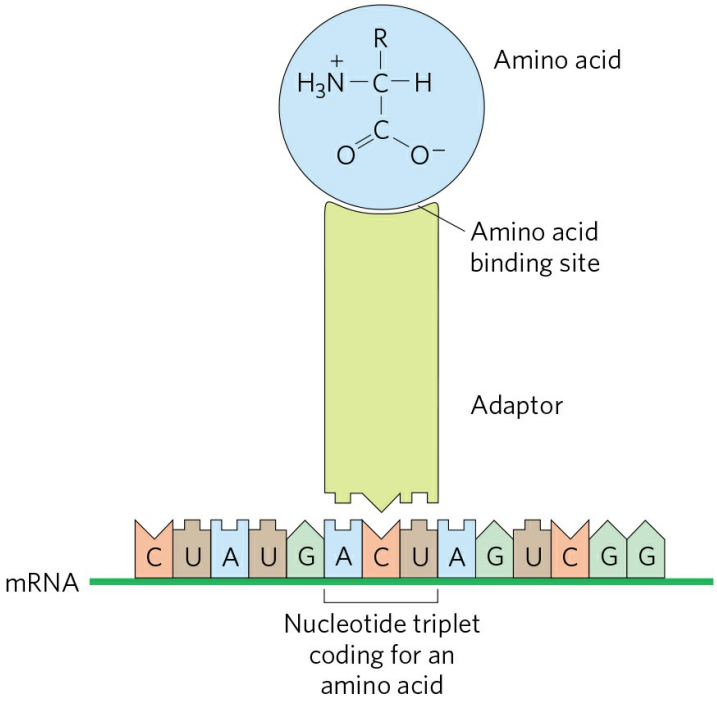


FIGURE 27-2 Crick's adaptor hypothesis. Today we know that the amino acid is covalently bound at the 3′ end of a tRNA molecule and that a specific nucleotide triplet elsewhere in the tRNA interacts with a particular triplet codon in mRNA through hydrogen bonding of complementary bases.

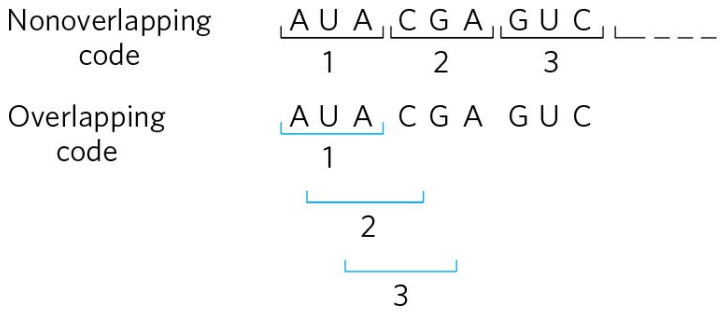


FIGURE 27-3 Overlapping versus nonoverlapping genetic codes. In a nonoverlapping code, codons (numbered consecutively) do not share nucleotides. In an overlapping code, some nucleotides in the mRNA are shared by different codons. In a triplet code with maximum overlap, many nucleotides, such as the third nucleotide from the left (A), are shared by three codons. Note that in an overlapping code, the triplet sequence of the first codon limits the possible sequences for the second codon. A nonoverlapping code provides much more flexibility in the triplet sequence of neighboring codons and therefore in the possible amino acid sequences designated by the code. The genetic code used in all living systems is now known to be nonoverlapping.

Several key properties of the genetic code were established in early genetic studies **(Figs 27-3, 27-4)**. A **codon** is a triplet of nucleotides that codes for a specific amino acid. Translation occurs in such a way that these nucleotide triplets are read in a successive, nonoverlapping fashion. A specific first codon in the sequence establishes the **reading frame**, in which a new codon begins every three nucleotide residues. There is no punctuation between codons for successive amino acid residues. The amino acid sequence of a protein is defined by a linear sequence of contiguous triplets. In principle, any given single-stranded DNA or mRNA sequence has three possible reading frames. Each reading frame gives a different sequence of codons **(Fig. 27-5)**, but only one is likely to encode a given protein. A key question remained: what were the three-letter code words for each amino acid?

In 1961, Marshall Nirenberg and Heinrich Matthaei reported the first breakthrough. They incubated synthetic polyuridylate, poly(U), with an *E. coli* extract, GTP, ATP, and a mixture of the 20 amino acids in 20 different tubes, each tube containing a different radioactively labeled amino acid. Because poly(U) mRNA is made up of many successive UUU triplets, it should promote the synthesis of a polypeptide containing only the amino acid encoded by UUU. A radioactive polypeptide was indeed formed in only one of the 20 tubes, the one containing radioactive phenylalanine. Nirenberg and Matthaei therefore concluded that the triplet codon UUU encodes phenylalanine. The same approach soon revealed that polycytidylate, poly(C), encodes a polypeptide containing only proline (polyproline), and polyadenylate, poly(A), encodes polylysine. Polyguanylate did not generate any polypeptide in this experiment because it spontaneously forms tetraplexes (see Fig. 8-20d) that cannot be bound by ribosomes.

Marshall Nirenberg, 1927–2010
[Source: AP Photo.]

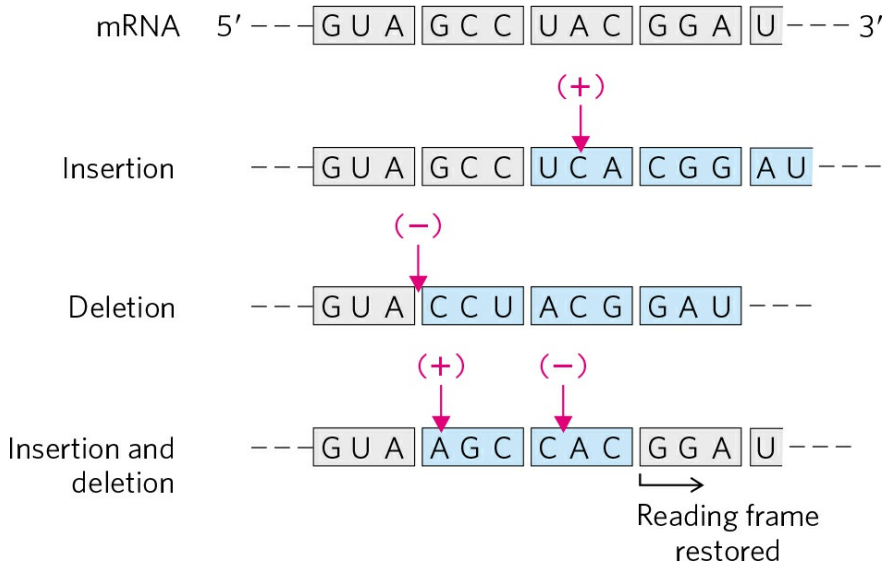


FIGURE 27-4 The triplet, nonoverlapping code. Evidence for the general nature of the genetic code came from many types of experiments, including genetic experiments on the effects of deletion and insertion mutations. Inserting or deleting one base pair (shown here in the mRNA transcript) alters the sequence of triplets in a nonoverlapping code; all amino acids coded by the mRNA following the change are affected. Combining insertion and deletion mutations affects some amino acids but can eventually restore the correct amino acid sequence. Adding or subtracting three nucleotides (not shown) leaves the remaining triplets intact, providing evidence that a codon has three, rather than four or five, nucleotides. The triplet codons shaded in gray are those

transcribed from the original gene; codons shaded in blue are new codons resulting from the insertion or deletion mutations.

The synthetic polynucleotides used in such experiments were prepared by using polynucleotide phosphorylase (p. 1063), which catalyzes the formation of RNA polymers starting from ADP, UDP, CDP, and GDP. This enzyme, discovered by Severo Ochoa, requires no template and makes polymers with a base composition that directly reflects the relative concentrations of the nucleoside 5′-diphosphate precursors in the medium. If polynucleotide phosphorylase is presented with UDP only, it makes only poly(U). If it is presented with a mixture of five parts ADP and one part CDP, it makes a polymer in which about five-sixths of the residues are adenylate and one-sixth are cytidylate. This random polymer is likely to have many triplets of the sequence AAA, smaller numbers of AAC, ACA, and CAA triplets, relatively few ACC, CCA, and CAC triplets, and very few CCC triplets (Table 27-1). Using a variety of artificial mRNAs made by polynucleotide phosphorylase from different starting mixtures of ADP, GDP, UDP, and CDP, the Nirenberg and Ochoa groups soon identified the base compositions of the triplets coding for almost all the amino acids. Although these experiments revealed the base composition of the coding triplets, they usually could not reveal the sequence of the bases.

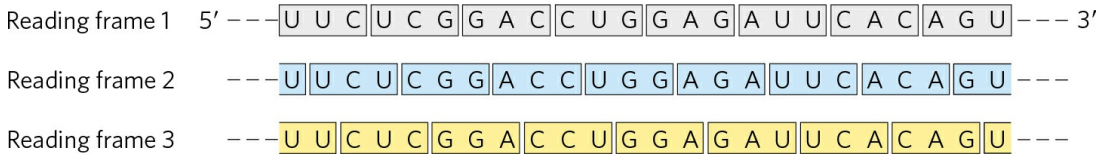


FIGURE 27-5 Reading frames in the genetic code. In a triplet, nonoverlapping code, all mRNAs have three potential reading frames, shaded here in different colors. The triplets, and hence the amino acids specified, are different in each reading frame.

TABLE 27-1 Incorporation of Amino Acids into Polypeptides in Response to Random Polymers of RNA

Observed	Tentative assignment for nucleotide	Expected frequency of incorporation

Amino acid	frequency of incorporation (Lys = 100)	composition of corresponding codon[a]	based on assignment (Lys = 100)
Asparagine	24	A_2C	20
Glutamine	24	A_2C	20
Histidine	6	AC_2	4
Lysine	100	AAA	100
Proline	7	AC_2, CCC	4.8
Threonine	26	A_2C, AC_2	24

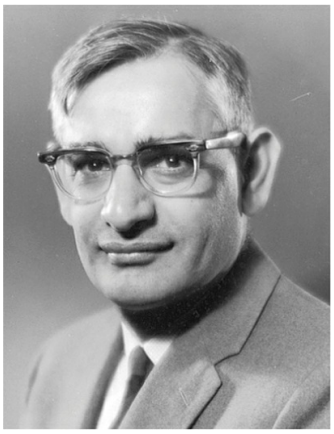
Note: Presented here is a summary of data from one of the early experiments designed to elucidate the genetic code. A synthetic RNA containing only A and C residues in 5:1 ratio directed polypeptide synthesis, and both the identity and the quantity of incorporated amino acids were determined. Based on the relative abundance of A and C residues in the synthetic RNA, and assigning the codon AAA (the most likely codon) a frequency of 100, there should be three different codons of composition A_2 C, each at a relative frequency of 20; three of composition AC_2 , each at a relative frequency of 4.0; and CCC at a relative frequency of 0.8. The CCC assignment was based on information derived from prior studies with poly (C). Where two tentative codon assignments are made, both are proposed to code for the same amino acid.

[a]These designations of nucleotide composition contain no information on nucleotide sequence (except, of course, AAA and CCC).

›› **Key Convention:** Much of the following discussion deals with tRNAs. The amino acid specified by a tRNA is indicated by a superscript, such as $tRNA^{Ala}$, and the aminoacylated tRNA by a hyphenated name: alanyl-$tRNA^{Ala}$ or Ala-$tRNA^{Ala}$. ‹‹

In 1964, Nirenberg and Philip Leder achieved another experimental breakthrough. Isolated *E. coli* ribosomes would bind a specific aminoacyl-tRNA in the presence of the corresponding synthetic polynucleotide messenger. For example, ribosomes incubated with poly(U) and phenylalanyl-$tRNA^{Phe}$ (Phe-$tRNA^{Phe}$) bind both RNAs, but if the ribosomes

are incubated with poly(U) and some other aminoacyl-tRNA, the aminoacyl-tRNA is not bound, because it does not recognize the UUU triplets in poly(U) (Table 27-2). Even trinucleotides could promote specific binding of appropriate tRNAs, so these experiments could be carried out with chemically synthesized small oligonucleotides. With this technique, researchers determined which aminoacyl-tRNA bound to 54 of the 64 possible triplet codons. For some codons, either no aminoacyl-tRNA or more than one would bind. Another method was needed to complete and confirm the entire genetic code.



H. Gobind Khorana, 1922–2011
[Source: Courtesy of Archives, University of Wisconsin–Madison.]

TABLE 27-2 Trinucleotides That Induce Specific Binding of Aminoacyl-tRNAs to Ribosomes

	Relative increase in ^{14}C-labeled aminoacyl-tRNA bound to ribosome[a]		
Trinucleotide	**Phe-tRNA^{Phe}**	**Lys-tRNA^{Lys}**	**Pro-tRNA^{Pro}**
UUU	4.6	0	0
AAA	0	7.7	0
CCC	0	0	3.1

Source: Information from M. Nirenberg and P. Leder, *Science* 145:1399, 1964.

[a]Each number represents the factor by which the amount of bound ^{14}C increased when the indicated trinucleotide was present, relative to a control with no trinucleotide.

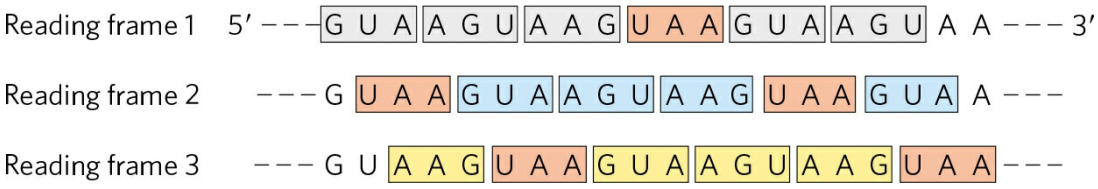


FIGURE 27-6 Effect of a termination codon in a repeating tetranucleotide. Termination codons (light red) are encountered every fourth codon in three different reading frames (shown in different colors). Dipeptides or tripeptides are synthesized, depending on where the ribosome initially binds.

At about this time, a complementary approach was provided by H. Gobind Khorana, who developed chemical methods to synthesize polyribonucleotides with defined, repeating sequences of two to four bases. The polypeptides produced by these mRNAs had one or a few amino acids in repeating patterns. These patterns, when combined with information from the random polymers used by Nirenberg and colleagues, permitted unambiguous codon assignments. The copolymer $(AC)_n$, for example, has alternating ACA and CAC codons: ACACACACACACACA. The polypeptide synthesized on this messenger contained equal amounts of threonine and histidine. Given that a histidine codon has one A and two Cs (Table 27-1), CAC must code for histidine and ACA for threonine.

Consolidation of the results from many experiments permitted assignment of 61 of the 64 possible codons. The other three were identified as termination codons, in part because they disrupted amino acid coding patterns when they occurred in a synthetic RNA polymer **(Fig. 27-6)**. Meanings for all the triplet codons (tabulated in Fig. 27-7) were established by 1966 and have been verified in many different ways.

The cracking of the genetic code is regarded as one of the most important scientific discoveries of the twentieth century.

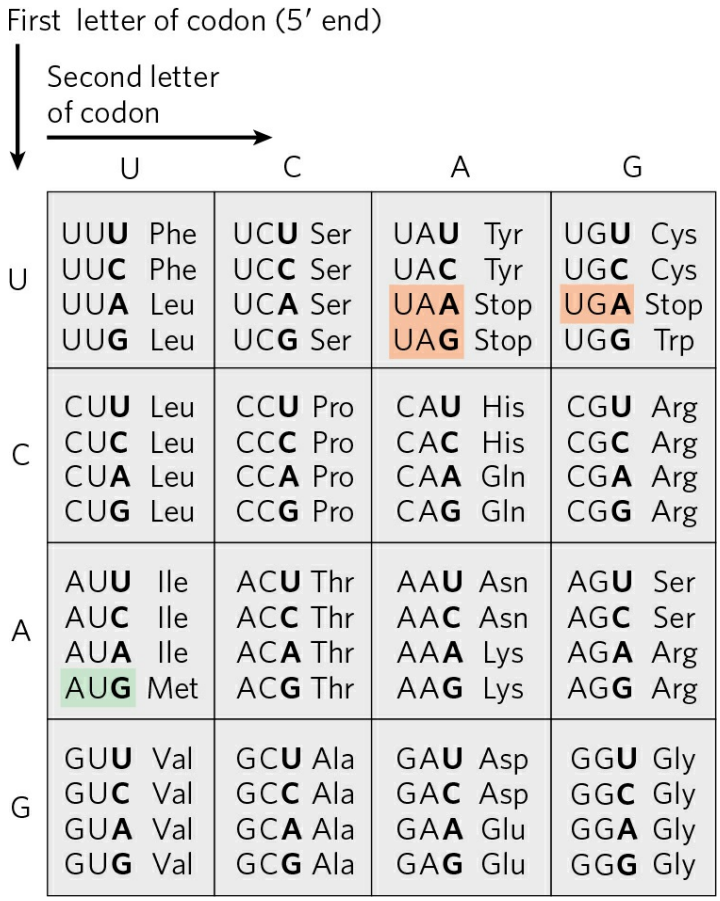


FIGURE 27-7 "Dictionary" of amino acid code words in mRNAs. The codons are written in the 5′→3′ direction. The third base of each codon (in bold type) plays a lesser role in specifying an amino acid than the first two. The three termination codons are shaded in light red, the initiation codon AUG in green. All the amino acids except methionine and tryptophan have more than one codon. In most cases, codons that specify the same amino acid differ only at the third base.