CONTENTS

CHAPTER

30

Genetic Code

Clusters of messenger RNA molecules (magnified 6700 times) form a fernlike structure around a backbone of DNA molecules, which are undergoing transcription. And it is on these mRNA molecules that the entire genetic dictionary is written.

GENERAL CONSIDERATIONS

As DNA is a genetic material, it carries genetic informations from cell to cell and from generation to generation. At this stage, an attempt will be made to determine that in what manner the genetic informations are existed in DNA molecule ? Are they written in articulated or coded language on DNA molecule? If in the language of codes, what is the nature of genetic code ?

A DNA molecule is composed of three kinds of moieties: (*i*) phosphoric acid, (*ii*) deoxyribose sugar, and (*iii*) nitrogen bases. The genetic informations may be written in any one of the three moieties of DNA. But the poly-sugar-phosphate backbone is always the same, and it is therefore unlikely that these moiteies of DNA molecule carry the genetic informations. The nitrogen bases, however, vary from one segment of DNA to another, so the informations might well depend on their sequences. *The sequences of nitrogen bases of a given segment of DNA molecule, actually has been found to be identical to linear sequence of amino acids in a protein molecule.* The proof of such a **colinearity** between DNA nitrogen base sequence and amino acid sequence in protein molecules was first obtained from an analysis of mutants of head protein of bacteriophage T_4 (**Sarabhai** *et al*, 1964) and the A protein of tryptophan synthetase of *Escherichia coli* (**Yanofski** *et al*, 1964). *The colinearity of protein molecules and DNA polynucleotides has given the clue that the specific arrangement of four nitrogen bases* (*e.g.*, *A, T, C and G*) *in DNA polynucleotide*

chains somehow determines the sequence of amino acids in protein molecules. Therefore, these four DNA bases can be considered as four alphabets of DNA molecule. All the genetic information, therefore, should be written by these four alphabets of DNA. Now the question arises that whether the genetic informations are written in articulated language or coded language ? If genetic informations might have occurred in an articulated language, the DNA molecule might require various alphabets, a complex system of grammer and ample amount of space on it. All of which might be practically impossible and troublesome too for the DNA. Therefore, it was safe to conclude for molecular biologists that genetic informations were existed in DNA molecule in the form of certain special language of code words which might utilize the four nitrogen bases of DNA for its symbols. Any coded message is commonly called **cryptogram**.

NATURE OF THE GENETIC CODE

Earlier, Gamow, the well-known nuclear physicist, proposed that the genetic code consists of three nitogenous (N) bases and the adjacent triplets overlap. This meant that at any particular point the same N-base occurs three times in a vertical manner instead of one which is expected on the basis of colinear model. This hypothesis, however, was not accepted on the following grounds :

1. In the overlapping model only certain amino acids can follow certain others. After the first amino acid in a protein is coded, the next two and for that matter the remaining amino acids in the protein are partially predetermined. If the first code is CAG, then the next must begin with AG and the third one with G.
2. Mutation involving a change in one base, according to this hypothesis, must involve three amino acids.

George Gamow (LT, 1904-1968), a Russian born US nuclear physicist and cosmologist, was one of the foremost advocates of the '*Big-bang theory*'. He is perhaps best known for his popular writings, designed to introduce to the nonspecialist such diffuse subjects as relativity and cosmology. His popular writings include : (1) *The creation of the Universe,* (2) *A Planet called Earth*, and (3) *A Star Called the Sun*. For his achievements as a popularizer of science, Gamow was awarded the **1956 Kalinga Prize** by UNESCO.

In order to find the **arrangement of codons**, in later experiments, it was found that when a change occurs due to a mutation, it is confined only to one amino acid. For instance, when sickle cell anemia occurs, only one amino acid, namely glutamic acid is changed into valine, the two adjacent amino acids remaining unaffected. Further research showed that the codons are arranged in a linear order. This explains as to why the change in one involves only one amino acid and not three; if Gamow's hypothesis were correct, change of one nitrogenous base would have involved 3 amino acids.

The sequence of bases that encodes a functional protein molecule is called a **gene.** And the **genetic code** is the relation between the base sequence of a gene and the amino acid sequence of the polypeptide whose synthesis the gene directs. In other words, the specific correspondence between a set of 3 bases and 1 of the 20 amino acids is called the genetic code.

J.D. Burke (1970) defined genetic code in the following words,

> "The genetic code for protein synthesis is contained in the base sequence of DNA. ... The genetic code is a code for amino acids. Specifically, it is concerned with what codons specify what amino acids."

The genetic code is the key that relates, in Crick's words, "...the two great polymer languages, the nucleic acid language and the protein language."

The "letters" in the "language" were found to be the bases; the "words" (codons) are groups of bases; and the "sentences" and "paragraphs" equate with groups of codons (*Eldon J. Gardner, 1968*).

Thus,

Letters ≡ Bases
Words ≡ Groups of bases (*i.e.*, codons)
Sentences and Paragraphs ≡ Groups of codons

The basic problem of such a genetic code is to indicate how information written in a four-letter-language (four nucleotides or nitrogen bases of DNA) can be translated into a twenty-letter-language (twenty amino acids of proteins). *The group of nucleotides that specifies one amino acid is a code word or codon.* The simplest possible code is a **singlet code** (a code of single letter) in which one nucleotide codes for one amino acid. Such a code is inadequate for only four amino acids could be specified. A **doublet code** (a code of two letters) is also inadequate because it could specify only sixteen (4×4) amino acids, whereas a **triplet code** (a code of three letters) could specify sixty four ($4 \times 4 \times 4$) amino acids. Therefore, it is likely that there may be 64 triplet codes for 20 amino acids. *The possible singlet, doublet and triplet codes, which are customarily represented in terms of "mRNA language", (mRNA is a complementary molecule which copies the genetic informations during its transcription*) can be illustrated as in Fig. 30–1. Larger than three letter units would seem wasteful and evidence already accumulated suggests that such larger units are unlikely.

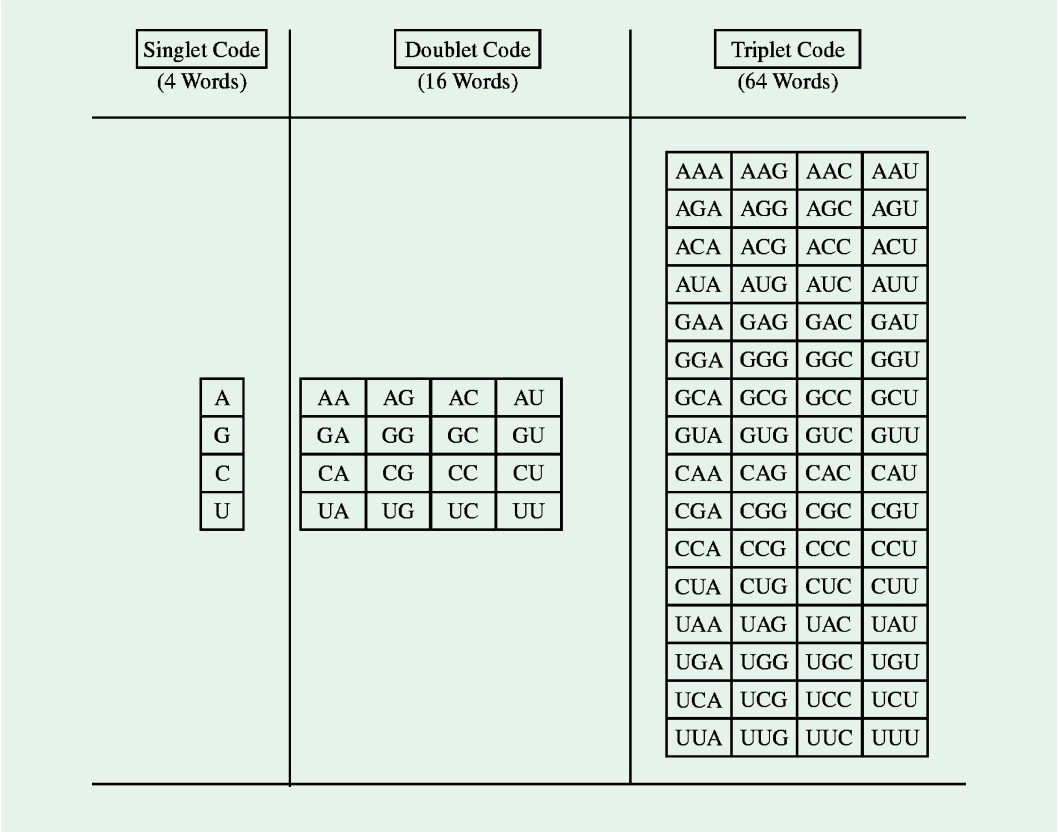


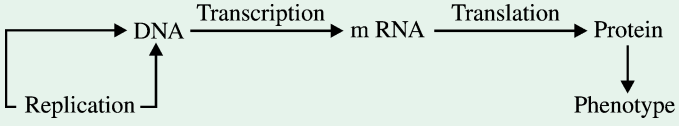
Fig. 30–1. Singlet, doublet and triplet codes of mRNA

The first experimental evidence in support to the concept of triplet codes is provided by **Crick** and coworkers in 1961. During their experiment, when they added or deleted single, or double base pairs in a particular region of DNA of T_4 bacteriophages of *E. coli*, they found that such bacteriophages ceased to perform their normal functions. However, bacteriophages with addition or deletion of three base pairs in DNA molecule, had performed normal functions. From this experiment, they concluded that *a genetic code is in triplet form because the addition of one or two nucleotides has put the reading of the code out of order*, while the addition of third nucleotide resulted in a return to the proper reading of the message.

A strong evidence in favour of triplet coding units is derived from determinations of the coding ratio. **Coding ratio** is equal to the number of nucleotides in the mRNA (or nucleotide pairs in the double-stranded DNA) divided by the number of amino acid residues of the resultant polypeptide chain. This expresses the number of nucleotides per coding unit. In each of the several genes so far studied, the number of nucleotide pairs of DNA has been estimated by genetic techniques and compared with the number of amino acid residues of the protein synthesised by the gene. All estimates give coding ratios close to three, indicating that three nucleotides compose a coding unit (triplet code). A wide variety of genetic experiments is consistent with a triplet code. The conclusion is inescapable that sequences of 3 bases in the mRNA molecule (triplet) contain coded information for the various amino acids. Such a triplet is called a **codon.** Each triplet codon specifies only one particular amino acid and the position of the codon in the mRNA molecule specifies the position of the amino acid in a polypeptide chain. As stated above, with four bases, 64 triplet codons (4^3) are possible. The genetic dictionary, thus, consists of 64 words, each made of a specific sequence of 3 out of 4 letters of the genetic alphabet. Any letter may occur more than once in a codon. *An example from the English language may be used to explain this*. Consider a 4-letter word "SEAT". Out of the four letters of this word, you can make many 3-letter words, each of which conveys a definite meaning, *e.g.*, "SEA", "SEE", "SET", "SAT", "EAT", "ASS", "ATE" and "TEA". The genetic code has now been experimentally deciphered and perfected by the combined efforts of many biochemists, notably Marshall Warren Nirenberg and Har Gobind Khorana, who were awarded the 1968 Noble Prize for their work, along with Robert Holley who was the first scientist to determine the nucleotide sequence of several tRNAs.

THE GENETIC CODE

The genetic language consists of only four letters contained in the word "GACU". These four letters can be combined to form 64 genetic words, each consisting of 3 letters. Each triplet word (codon) has a specific meaning which the cell understands. It codes for a particular amino acid. The genetic code, as at present known, is shown in Fig. 30–2. It shows the base sequences of the various codons (triplets) and against each codon is given the amino acid that it codes. Just as different combinations of different words make different sentences, each having a specific meaning, similarly different sequences of codons on mRNA specify different proteins, each with a specific sequence of amino acids. Although the genetic information resides in DNA, the terms genetic code and codon are used with reference to mRNA because mRNA is the nucleic acid which directly determines the sequence of amino acids in a protein. This expression of genetic information in the amino acid sequence of proteins by mRNA is called **translation.** *The DNA-RNA-Protein code may be expressed as under :*



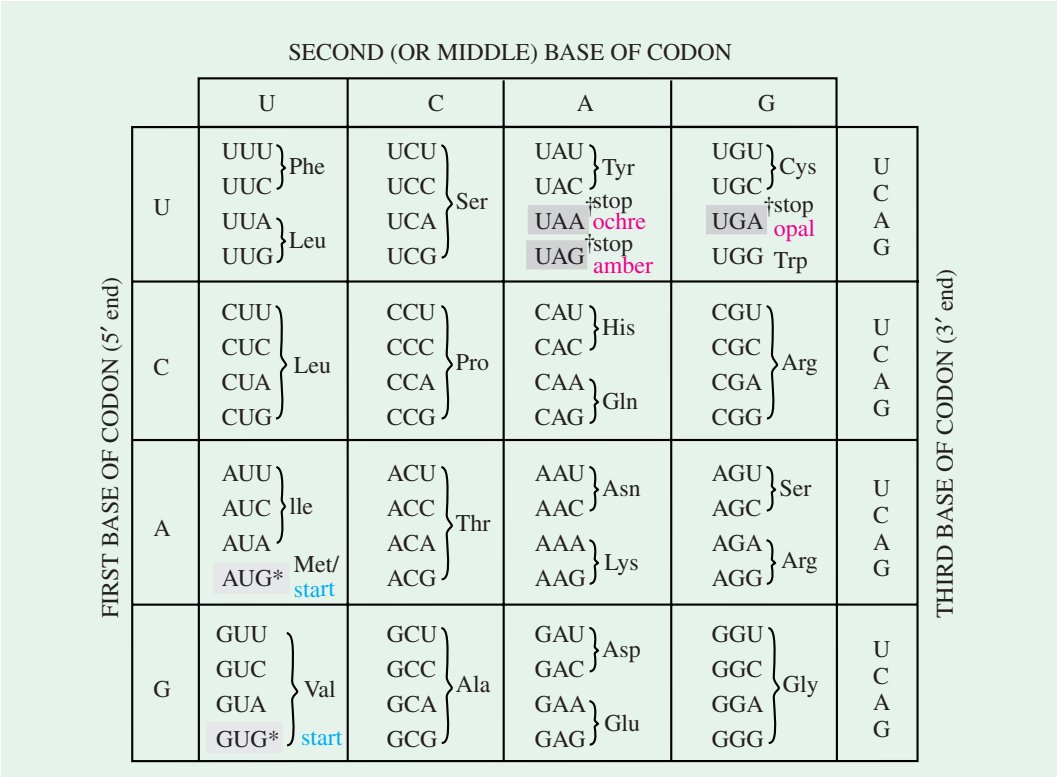


Fig. 30–2. The genetic code dictionary

The 64 triplet codons are listed in the 5′ ⟶ 3′ direction in which they are read. Against each condon is written the name of the amino acid encoded by it. The first genetic code ever to be established was the codon (UUU) for phenylalanine (Phe). AUG and GUG encode methionine, which initiates most polypeptide chains. All other amino acids, except tryptophan (which is encoded only by UGG), are represented by 2 to 6 triplets. The 3 triplets UAA, UAG and UGA are termination signals and do not encode any amino acids. UGA rarely codes for selenocysteine (SeCys).

To read code dictionary : A codon consists of three nucleotides read in the sequence indicated by the column heads. For example, ACU codes threonine. The first letter, A, is read in the first-letter column; the second letter, C, from the second-letter column; and the third letter, U, from the third-letter column. Each codon is recognized by a corresponding anticodon sequence on a tRNA molecule. Some tRNA molecules recognize more than one codon sequence but always for the same amino acid. Most amino acids are encoded by more than one codon. For example, threonine is encoded by four codons (ACU, ACC, ACA, and ACG), which differ from one another only in the third position.

* Chain initiation codons

† Chain termination codons

The synthesis of cellular proteins takes place in the joining together of several amino acids to form a linear polypeptide chain of variable length. There are 20 different amino acids which are commonly found in protenis (hence called protein amino acids) and which take part in their synthesis. The mRNA codons for these 20 protein amino acids, as can be deduced from Fig. 30–2, are given in Fig. 30–3.

	Amino Acid	Abbreviation	mRNA Codons		Total no. of codon(s)
			Common bases	Complete codon(s)	
1.	Alanine	Ala	GC–	GCU, GCC, GCA, GCG	4
2.	Arginine	Arg	CG– AG–	CGU, CGC, CGA, CGG AGA, AGG	6
3.	Asparagine	Asn	AA–	AAU, AAC	2
4.	Aspartic acid	Asp	GA–	GAU, GAC	2
5.	Cysteine	Cys	UG–	UGU, UGC	2
6.	Glutamic acid	Glu	GA–	GAA, GAG	2
7.	Glutamine	Gln	CA–	CAA, CAG	2
8.	Glycine	Gly	GG–	GGU, GGC, GGA, GGG	4
9.	Histidine	His	CA–	CAU, CAC	2
10.	Isoleucine	Ile	AU–	AUU, AUC, AUA	3
11.	Leucine	Leu	UU– CU–	UUA, UUG CUU, CUC, CUA, CUG	6
12.	Lysine	Lys	AA–	AAA, AAG	2
13.	Methionine	Met	AU–	AUG	1
14.	Phenylalanine	Phe	UU–	UUU, UUC	2
15.	Proline	Pro	CC–	CCU, CCC, CCA, CCG	4
16.	Serine	Ser	UC– AG–	UCU, UCC, UCA, UCG AGU, AGC	6
17.	Threonine	Thr	AC–	ACU, ACC, ACA, ACG	4
18.	Tryptophan	Trp	UG–	UGG	1
19.	Tyrosine	Tyr	UA–	UAU, UAC	2
20.	Valine	Val	GU–	GUU, GUC, GUA, GUG	4
Terminator triplets		**Trm**	UA– UG–	UAA, UAG UGA	3
				Total	**64**

Fig. 30–3. Proteinogenic amino acids and their mRNA codons

CHARACTERISTICS OF THE GENETIC CODE

The genetic code is endowed with many characteristic properties which have actually been proved by definite experimental evidences. These are described below :

1. Triplet nature

As earlier outlined, singlet and doublet codes are not adequate to code for 20 amino acids; therefore, it was pointed out that triplet code is the minimum required. But it could be a quadruplet code or of a higher order. As pointed out above, in a triplet code of 64 codons, there is an excess of (64 – 20) = 44 codons and, therefore, more than one codons are present for the same amino acid. This excess will be still greater if more than three-letter words are used. In a *quadruplet code* there will be 4^4 (4 × 4 × 4 × 4) = 256 possible codons. An account of the 20 amino acids along with their corresponding codons is presented below :

2	amino acids (Met, Trp)	... have 1 codon each	=	2
9	amino acids (Asn, Asp, Cys, Gln, Glu, His, Lys, Phe, Tyr)	... have 2 codons each	=	18
1	amino acid (Ile)	... has 3 codons	=	3
5	amino acids (Ala, Gly, Pro, Thr, Val)	... have 4 codons each	=	20
3	amino acids (Arg, Leu, Ser)	... have 6 codons each	=	18
		3 terminator codons	=	3
20	Amino acids			64

A closer scrutiny of the Genetic Dictionary (Fig. 30–2) has revealed the emergence of certain trends for **patterns of the genetic code.**

1. Amino acids with similar structural properties tend to have related codons. Thus, aspartic acid codons (GAU, GAC) are similar to glutamic acid codons (GAA, GAG); the difference being exhibited only in the third base (toward 3′ end).
2. Similarly, the codons for the aromatic amino acids phenylalanine (UUU, UUC), tyrosine (UAU, UAC) and tryptophan (UGG) all begin with uracil (U).
3. The first two bases of all the 4 codons assigned to each of the 5 amino acids are similar : GC for alanine, GG for glycine, CC for proline, AC for threonine and GU for valine.
4. All codons with U in the second position specify hydrophobic amino acids (Ile, Leu, Met, Phe, Val).
5. All codons with A in the second position specify the charged amino acids, except Arg.
6. All the acidic (Asp, Glu) and basic (Arg, Lys) amino acids have A or G as the second base.

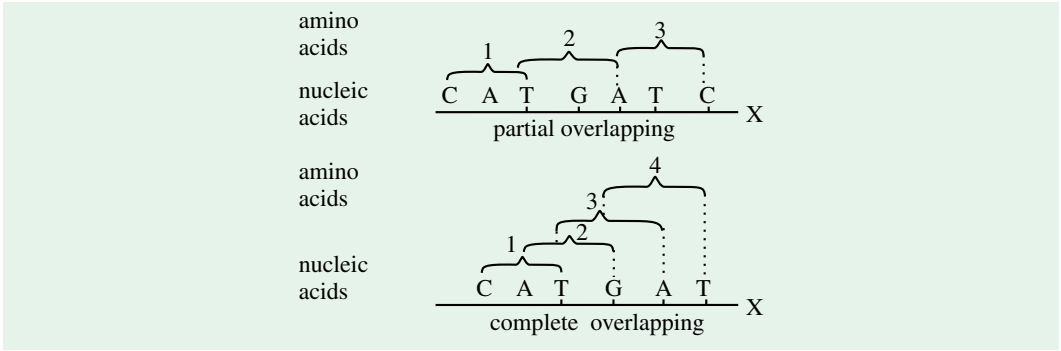


Fig. 30–4. Overlapping of codons due to one letter or two letters

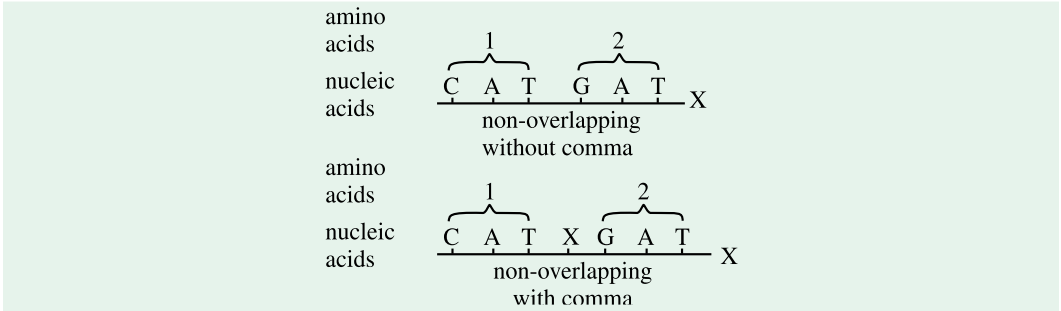


Fig. 30–5. Genetic code, without comma and with comma

2. Degeneracy

The code is degenerate which means that the same amino acid is coded by more than one base triplet. Degeneracy, as used here, does not imply lack of specificity in protein synthesis. It merely

means that a particular amino acid can be directed to its place in the peptide chain by more than one base triplets (Eldon J. Gardner, 1968). For example, the three amino acids arginine, alanine and leucine each have six synonymous codons. A non-degenerate code would be one where there is one to one relationship between amino acids and the codons, so that from the 64 codons, 44 will be useless or nonsense codons. It has been definitely shown that there are no nonsense codons. The codons which were initially called nonsense codons were later shown to mean stop signals.

The code degeneracy is basically of 2 types: partial and complete. In *partial degeneracy*, the first two nucleotides are identical but the third (*i.e.,* 3′ base) nucleotide of the degenerate codon differs; for example, CUU and CUC code for leucine. *Complete degeneracy* occurs when any of the 4 bases can take third position and still code for the same amino acid; for example, UCU, UCC, UCA and UCG all code for serine.

Degeneracy of genetic code has certain biological advantages. For example, it permits essentially the same complement of enzymes and other proteins to be specified by the microorganisms varying widely in their DNA base composition. Degeneracy also provides a mechanism of minimizing mutational lethality.

Degeneracies occur frequently in the third letter of the codon. Exceptions are, however, arginine (Arg), leucine (Leu) and serine (Ser) which have 2 groups of codons or triplets, which differ in either the first base only (Arg, Leu) or in both the first and second bases (Ser).

3. Nonoverlapping

The genetic code is nonoverlapping, *i.e.*,the adjacent codons do not overlap. A nonoverlapping code means that the same letter is not used for two different codons. In other words, no single base can take part in the formation of more than one codon. Fig. 30–4 shows that an overlapping code can mean coding for four amino acids from six bases. In actual practice, six bases code for not more than two amino acids. As an illustration, an end-to-end sequence of 5′ UUUCCC 3′ on mRNA will code only 2 amino acids, *i.e.,* phenylalanine (UUU) and proline (CCC).

4. Commaless

There is no signal to indicate the end of one codon and the beginning of the next. The genetic code is commaless (or comma-free). A commaless code means that no codon is reserved for punctuations or the code is without spacers or space words. There are no intermediary nucleotides (or commas) between the codons. In other words, we can say that after one amino acid is coded, the second amino acid will be automatically coded by the next three letters and that no letters are wasted for telling that one amino acid has been coded and that second should now be coded (Fig. 30–5).

5. Non-ambiguity

By non-ambiguous code, we mean that there is no ambiguity about a particular codon. A particular codon will always code for the same amino acid. In an ambiguous code, the same codon could code for two or more than two different amino acids. Such is not the case. *While the same amino acid can be coded by more than one codon (the code is degenerate), the same codon shall not code for two or more different amino acids (non-ambiguous).* But sometimes the genetic code is ambiguous, that is, same codon may specify more than one amino acid. For example, UUU codon usually codes for phenylalanine but in the presence of streptomycin, may also code for isoleucine, leucine or serine.

6. Universality

The genetic code applies to all modern organisms with only very minor exceptions. Although the code is based on work conducted on the bacterium *Escherichia coli* but it is valid for other organisms. This important characteristic of the genetic code is called its **universality.** It means that the same sequences of 3 bases encode the same amino acids in all life forms from simple microorganisms to complex, multicelled organisms such as human beings. Consider any codon. It codes for the same amino acid from the smallest organism to the largest, plant or animal. Thus, UUU codes for

phenylalanine and GUC for valine in all living things, from amoeba to ape, bacteria to the banyan tree, and from cabbage to kings. The genetic code which was first developed in the bacteria about 3 billion (300 crore) years ago has not undergone any change and has been preserved in its almost original form in the course of evolution. In other words, *the code is a conservative one*, *i.e.*, the code was fixed early in the course of evolution and has been maintained to the present day.

7. Polarity

The genetic code has polarity, that is, the code is always read in a fixed direction, *i.e.*, in the $5' \rightarrow 3'$ direction. It is apparent that if the code is read in opposite direction (*i.e.*, $3' \rightarrow 5'$), it would specify 2 different proteins, since the codon would have reversed base sequence :

Codon	:	UUG	AUC	GUC	UCG	CCA	ACA	AGG
Polypeptide	:	$\rightarrow$ Leu	Ile	Val	Ser	Pro	Thr	Arg
		Val	Leu	Leu	Ala	Thr	Thr	Gly $\leftarrow$

8. Chain Initiation Codons

The triplets AUG and GUG play double roles in *E. coli*. When they occur in between the two ends of a cistron (intermediate position), they code for the amino acids methionine and valine, respectively in an intermediate position in the protein molecule. But when they occur immediately after a terminator codon, they act as "chain initiation" (C.I.) signals or "starter codons" for the synthesis of a polypeptide chain. It has also been shown that the initiating methionine molecule should be found in the formylated state. This makes a distinction between the initiating methionine and the methionine at internal position. The methionine when required at internal position should not be formylated. Also while formyl methionine is carried by $tRNA^{fMet}$, there is a separate species of tRNA for internal methionine and it is designated as $tRNA^{mMet}$.

9. Chain Termination Codons

The 3 triplets UAA, UAG, UGA do not code for any amino acid. They were originally described as **non-sense codons,** as against the remaining 61 codons, which are termed as **sense codons.** The so-called non-sense codons have now been found to be of "special sense". When any one of them occurs immediately before the triplet AUG or GUG, it causes the release of the polypeptide chain from the ribosome. Hence, the use of the term 'non-sense' is unfortunate. These *special-sense codons* perform the function of punctuating genetic message like a full stop at the end of a sentence. They are also called chain termination codons because these codons are used by the cell to signal the natural end of translation of a particular peptidyl chain. However, their inclusion in any mRNA results in the abrupt termination of the message at the point of their location even though the polypeptide chain has not been completed. The codons UAA and UAG were discovered in bacteria and were respectively associated with the *ochre* and *amber* mutations. Hence, UAA is also called **ochre** and UAG is also known as **amber** (because an investigator who studied the properties of this codon belonged to the Bernstein family, and Bernstein means amber in German). UGA is also called **opal.** They resulted in the formation of incomplete polypeptide chains. UGA is the usual terminator codon in all cases.

Most of these principles are illustrated by the following analogy. Consider a sentence (gene) in which the words (codons) each consist of 3 letters (bases).

THE BIG RED FOX ATE THE EGG

(Here the spaces separating the words have no physical significance; they are only present to indicate the reading frame.) The deletion of the 4th letter, which shifts the reading frame, changes the sentence to

THE IGR EDF OXA TET HEE GG

so that all words past the point of deletion are unintelligible (specify the wrong amino acids). An insertion of any letter, however, say an X in the 9th position,

THE IGR EDX FOX ATE THE EGG

restores the original reading frame. Consequently, only the words between the two changes (mutations) are altered. As in this example, such a sentence might still be intelligible (the gene could still specify a functional protein), particularly if the changes are close together. Two deletions or two insertions, no matter how close together, would not suppress each other but just shift the reading frame. However, three insertions, say X, Y, and Z in the 5th, 8th and 12th positions, respectively, would change the sentence to

THE BXI GYR EDZ FOX ATE THE EGG

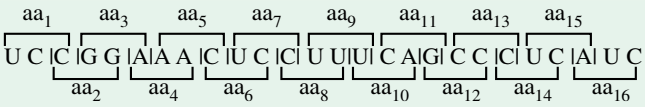
which, after the third insertion, restores the original reading frame. The same would be true of three deletions. As before, if all three changes were close together, the sentence might still retain its meaning.

Crick and Brenner did not unambiguously demonstrate that the genetic code is a triplet code because they had no proof that their insertions and deletions involved only single nucleotides. Strictly speaking, they showed that a codon consists of $3r$ nucleotides where r is the number of nucleotides in an insertion or deletion. Although it was generally assumed at the time that $r = 1$, proof of this assertion had to await the elucidation of the genetic code.

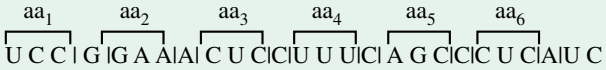
DECIPHERING THE GENETIC CODE OR CODON ASSIGNMENT

The early studies that led to the breaking of the genetic code are serendipitous. Following the discovery of mRNA, the immediate question was: How is it read? It was clear that a triplet code was the minimum required to account for all 20 amino acids, but many kinds of triplet codes are possible. Some conceivable ones are shown in Fig. 30–6. An *overlapping code* like that shown in Fig. 30–6(*a*) would be space-saving. This possibility was eliminated, however, by the observation that most mutations result in the change of a single amino acid. If the codons overlapped, there should be a significant number of cases in which two adjacent residues are modified by mutation of the "overlap" residue. Furthermore, an overlapping code would lead to statistical regularities between neighbouring amino acid residues in proteins—that is, some amino acids would be neighbours more often than others—and this has never been observed. Another possibility was that the code

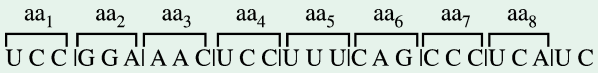
Serendipitous is adjective of the word serendipity which means occurrence of events by chance in a fortunate way. The word **'serendipity'** was coined ca 1754 by Horace Walpole after *The Three Princes of Serendip* (Serendip, a former name for Sri Lanka), a Pers fairy tale in which the princes make such discoveries.



(*a*) **Overlapping code :** There will be statistical regularities between adjacent amino acid residues. Point mutations will be able to change two amino acid residues.



(*b*) **Punctuated code :** Deletions of four nucleotides (or multiples thereof) will restore the reading frame.



(*c*) **Unpunctuated code :** Deletions of three nucleotides (or multiples thereof) will restore the reading frame. This is the actual form of the code.

Fig. 30–6. Three conceivable types of genetic codes

Early research on the nature of the code quickly showed that a nonoverlapping, unpunctuated code (*c*) fit all experimental observations.

was *punctuated*. That is, as shown in Fig. 30–6 (*b*), some base or bases might serve as "spacing" between code words.

We may here briefly describe the techniques which ultimately resulted in determining the code words for each of 20 proteinogenic amino acids. Most of our current knowledge of the general nature of the genetic code and the nucleotide composition were obtained from the following *four main types of experimental approaches* :

A. Assignment of Codons with Unknown Sequences

1. Polyuridylic Acid Method : This method for establishing the genetic code consists of acids in the protein molecule synthesized under its influence. The sequence analysis for proteins is now a standard biochemical technique but unfortunately the determination of the base sequence of large nucleic acid molecules (like mRNA) is yet to be achieved. Biochemists, therefore, adopted indirect methods to crack the genetic code. A cell-free amino acid incorporating system and the polymerizing enzyme were known which could polymerize single ribonucleoside triphosphates or their mixtures to produce long nucleotide chains of random base sequence. *Initial breakthrough in this direction was achieved by* **Marshall Nirenberg and Heinrich Mathei** *in 1961 who adopted the following technique.* Using the enzyme *polynucleotide phosphorylase*, they prepared a synthetic polyribonucleotide containing only one kind of base and added it to a cell-free amino acid-incorporating system from *E. coli*. In this way, they prepared, for example, a polyuracil (UUU...UUU) to which they added different amino acids labelled with C^{14}. They found that this poly-U system specifically stimulated the formation of a polypeptide which contained only the amino acid *phenylalanine*. The conclusion was drawn that the code word for the amino acid phenylalanine was a sequence of three uracil nucleotides (UUU) in mRNA. This was the first demonstration that synthetic polyribonucleotide could act as a messenger for polypeptide synthesis. Later on, **Nirenberg and Severo Ochoa** became deeply involved in this problem. Subsequently, they found that poly A gave polylysine and poly C gave polyproline. Therefore, AAA was assigned to *lysine* and CCC to *proline*. Similar experiments with poly G were not successful, because it attains secondary structure and thus cannot attach to ribosomes. The following three codons thus could be assigned, using homopolymers of RNA, without any difficulty : UUU = *phenylalanine;* AAA = *lysine;* CCC = *proline.*

2. Copolymers Method : In this method, Nirenberg used mixtures of two or more ribonucleoside disphosphates and with the help of the enzyme mentioned above, polyribonucleotides were prepared. Thus, using UDP and CDP in the ratio of 3:1, he obtained a polynucleotide which contained the following triplets in the descending order of frequency: UUU, UUC, UCU and CUU. The triplets containing 2Cs and IU were the least frequent. With such a poly-UC system, Nirenberg obtained a

SEVERO OCHOA
(LT, 1905–1993)

Severo Ochoa was educated in his home country, graduating from Malaga University in 1921 and then obtaining his MD from Madrid in 1929. He worked in Germany with Otto Meyerhof on muscle biochemistry and then moved via England to the USA, joining New York University in 1942. During the 1950s his research centred on the enzymes that utilize the energy held in the energy-rich phosphate bonds of ATP. This led in 1955 to his discovery, with Marianne Grunberg-Manago, of polynucleotide phosphorylase, the enzyme that was subsequently used to prepare synthetic mRNA molecules for elucidation of the genetic code. Ochoa received the **1959 Nobel Prize for Physiology or Medicine**. In 1985 he returned to Madrid University as Professor of Biology.

His advice to the science students regarding conductions of research reads as follows :

" My advice to students of science is that if they have an urge to do research they should to it by all means. Nothing should stand in the way of a strong wish to devote Life to Science."

" If you have the urge to do scientific research get the proper training and by all means do it; nothing else is likely to give you so much satisfaction and, above all such a sense of fulfilment."

polypeptide containing the amino acids phenyalanine and serine in the ratio 3:1. He concluded that the code word for serine contains 2Us and 1C. This method, however, did not give the exact sequence of the three bases. Ochoa also used this technique and determined the code words for most of the amino acids.The exact sequence of the three bases for any amino acid was not known because of these polynucleotides had random base sequence. The assigning of the exact triplet for every amino acid was not possible at this stage.

In an experiment, Nirenberg and his coworkers used RNA synthesized by using two or more bases. For instance, if only A and C are used, poly AC will consist of eight possible codons, namely AAA, AAC, ACA, CAA, CCA, CAC, ACC and CCC. The proportion of these eight codons, in the synthetic RNA can be calculated if the known quantities of A and C are used for the synthesis of poly AC. For instance, if A : C = 5 : 1, (5/6 is A and 1/6 is C), the calculated relative proportions of eight codons on random basis would be as given in Table 30–1.

Table 30–1. The relative proportions of different codons in mRNA formed due to bases A : C taken in 5 : 1 ratio.

Base composition	*Codon and probability*	*Ratio using*	
		minimum (as 1)	*maximum (as 100)*
(1) 3A	1. AAA 5/6 × 5/6 × 5/6 = 125/216	125	100
(2) 2A1C	2. AAC 5/6 × 5/6 × 1/6 = 25/216	25	20
	3. ACA 5/6 × 1/6 × 5/6 = 25/216	25	20
	4. CAA 1/6 × 5/6 × 5/6 = 25/216	25	20
(3) 1A2C	5. CCA 1/6 × 1/6 × 5/6 = 5/216	5	4
	6. CAC 1/6 × 5/6 × 1/6 = 5/216	5	4
	7. ACC 5/6 × 1/6 × 1/6 = 5/216	5	4
(4) 3C	8. CCC 1/6 × 1/6 × 1/6 = 1/216	1	0.80

The calculated relative proportions of codons were compared with the proportions in which different amino acids were present in the polypeptides synthesized using poly AC. For instance, if an amino acid is 1/5th of lysine (coded by AAA), we can say that it should be coded by one of the three possible 2A1C codons (AAC, ACA or CAA). Similar reasoning would allow assignments of 1A2C as well as 3C. However, using this technique, it was not possible to assign the three codons of the category 2A1C (*i.e.*, AAC, ACA, CAA) to three amino acids, since these will be present in equal quantities. Therefore, the codons were initially assigned only with respect to base composition, ignoring the sequences of the bases in codons, as done in the above example. The assignments are given in Table 30–2.

Table 30–2. Codon assignments derived due to use of A : C (5 : 1) in the synthesis of mRNA

Amino acids	*Codon composition*
(1) lysine	3A
(2) asparagine, glutamine and threonine	2A1C
(3) histidine, proline and threonine	1A2C
(4) proline	3C

B. Assignment of Codons with Known Sequences

1. Binding Technique : Marshall W. Nirenberg and **Philip Leder** in 1964 found that if a synthetic trinucleotide for a known sequence (with known bases at 5' end and 3' end) is used with ribosome and a particular aminoacyl-tRNA (tRNA having its own specific amino acid attached), these will form a complex, provided the used codon codes for the amino acid attached to the given aminoacyl tRNA.

$Codon_1 + Ribosome + AA_1 + tRNA \rightarrow Ribosome\text{-}Codon_1\text{-}AA_1\text{-}tRNA_1$

In a process such as above, if given AA_1 is used with a given codon, and the formation of the complex is detected, this would prove that the given codon codes for the given amino acid.

MARSHALL NIRENBERG

It was also observed that while the free AA-tRNA passed through nitrocellulose membrane, the **ribosome-codon-AA-tRNA complex** adsorbs on such a membrane. If in a particular mixture only one of the amino acids is made radioactive, then the presence or absence of the radioactivity on the nitrocellulose membrane will show whether there is a relationship between the codon and the amino acid which was made radioactive. For instance, 20 samples of a mixture of all 20 amino acids may be taken and in each sample one amino acid is made radioactive in such a manner that each and every amino acid is made radioactive in one sample or the other, and no two samples have same radioactive amino acid. A particular sample would be then known by its radioactive amino acid. Now tRNAs and ribosomes are mixed with each sample and same codon is used for complex formation in all 20 cases. When the mixture is poured on the nitrocellulose membrane, radioactivity on membrance will be observed only when the radioactive amino acid is taking part in the formation of the complex. Since in each sample the radioactive amino acid is known, it would be possible to detect the amino acid coded by a given codon by the presence of radioactivity on the membrane. Such a treatment was given by Nirenberg and his coworkers to all the 64 synthetic codons, and their respective amino acids were identified. The binding of AA-tRNA was not equally efficient in all cases. Therefore, the sequences of bases in only about 45 codons could be worked out by this method.

HAR GOBIND KHORANA

(Born, 1922 in India) Khorana received a bachelor's and a master's degree from Punjab University and a Ph.D. from the University of Liverpool. In 1960 he joined the faculty at the University of Wisconsin and later became a professor at MIT.

2. Repetitive Sequencing Technique: This method of confirming the genetic code is the most direct method and was devised by the Nobel prize winner **Prof. Har Gobind Khorana.** This method consists of *in vitro* chemical synthesis of short segments of DNA of known base sequence with the help of DNA polymerase. From this synthetic DNA, a polyribonucleotide (RNA) of strictly defined base sequence is transcribed under the catalytic influence of RNA polymerase. *A polypeptide is then synthesised under the direction of RNA as represented in* Fig. 30–7.

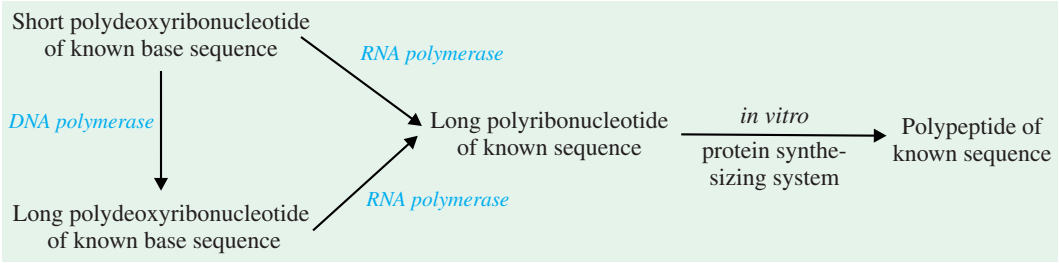


Fig. 30–7. Synthesis of polypeptide under the direction of RNA

The amino acid sequence of the polypeptide so formed is then determined and correlated with the base sequence of DNA and RNA.

Using homopolynucleotides (all bases the same) as templates could yield only a few code words. Many more words were deciphered after Khorana developed methods for synthesizing polyribonucleotides with different but repeating structures. In the example shown in Fig. 30–8, the repeating sequence $(AAG)_n$ was found to give 3 different homopolymers : polylysine, polyarginine and polyglutamic acid. This finding not only confirmed the importance of the reading frame but also showed that AAG, AGA and GAA must be codons for these amino acids. However, the experiment

did not reveal which codon corresponded to which amino acid; further experiments were needed to discriminate among the possible matches.

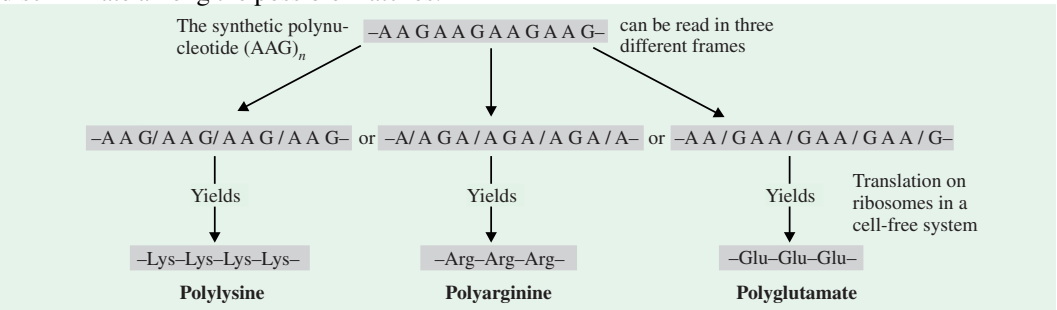


Fig. 30–8. Use of synthetic polynucleotides with repeating sequences to decipher the code
This example shows how polypeptides derived from the $(AAG)_n$ polymer were used to confirm the triplet code and help identify codons. The polymer $(AAG)_n$ can yield 3 different polypeptides, depending on which reading frame is employed.

Using synthetic DNA, Khorana and his coworkers could prepare polyribonucleotides (RNA) with known repeating sequences. A repeating sequence means that, if CU are two bases, these will be repeatedly present throughout the length as follows :

CUCUCUCUCUCU

In a similar manner, if ACU are three bases, they will be present repeatedly as follows :

ACUACUACUACU

Such copolymers will direct the incorporation of amino acids in a manner which can be theoretically predicted. For instance, in $(CU)_n$ =(CUC/UCU/CUC/UCU), only two codons are possible and these are CUC and UCU. Moreover, these codons are present in alternating sequence. The result would be that the polypeptide formed would have only two amino acids in alternating sequences. These two amino acids can be assigned to the two codons (Table 30–3).

Table 30–3 Assignment of codons, having known sequences, with the help of copolymers having repetitive sequences of two bases.

Copolymers	*Codons*	*Amino acids*	*Codons*
$(CU)_n$	CUC/UCU/CUC	leucine/serine	CUC/UCU
$(UG)_n$	UGU/GUG/UGU	cysteine/valine	UGU/GUG
$(AC)_n$	ACA/CAC/ACA	threonine/histidine	ACA/CAC

We may similarly consider a repeating sequence of three bases *e.g.,* $(ACG)_n$. Depending upon where the reading is started, three kinds of homopolypeptides are expected (Table 30–4). Actual codon assignment *i.e,.* to find out which of the three codons codes for which amino acid would depend upon the previous information available regarding the composition of bases in different codons coding for different amino acids.

Table 30–4. Assignment of codons, having known sequences, with the help or copolymers having repetitive sequeuces of three bases = $(ACG)_n$.

Codons	*Homopolypeptide*	*Codon assignment*
ACG/ACG/ACG/ACG/ACG = Poly (ACG)	(threonine)n	ACG = threonine
A/CGA/CGA/CGA/CGA/CGA = Poly (CGA)	(arginine)n	CGA = arginine
AC/GAC/GAC/GAC/GAC = poly (GAC)	(aspartic acid)n	GAC = aspartic acid

These studies of Khorana and his coworkers with chemically-defined messengers proved very conclusively that (1) the base sequence in DNA specifies the sequence of amino acids in proteins, (2)

the information contained in DNA is conveyed through RNA, and (3) the genetic code is triplet and non overlapping in nature. The results obtained from these polypeptide synthesis in combination with the results obtained from the binding technique of Nirenberg and coworkers led to the deciphering and establishment of the genetic code.

By this technique, Khorana not only worked out and confirmed the exact sequence of code words for all the amino acids but also clarified the roles of the 5′ and 3′ terminals of mRNA molecule. He showed that the translation of mRNA proceeded from its 5′-hydroxyl end towards its 3′-hydroxyl end. His investigations also led to the artificial production of small segments of DNA molecule, thereby paving the way for the synthesis of artificial genes which could function in a living cell. In 1970, a Japanese student of Khorana, **Miss Otsuka**, succeeded in chemically linking six nucleotides by attaching an extra phosphoric acid artificially to a nucleotide in addition to the one it already had. By this process, Otsuka and her coworkers at the Osaka University synthesized alanine-tRNA. This synthetic tRNA had exactly the same sequence of nucleotides as the one that occurs in natural cells.

On the basis of the above techniques, a complete genetic code dictionary could be prepared which has been presented in Fig. 30–2.

MULTIPLE RECOGNITION OF CODONS AND WOBBLE HYPOTHESIS

It has been observed that change in the first base of a codon requires a new tRNA species for its recognition. *In most cases, the first two bases for the possible triplets for each amino acid are always the same.* The third base can vary (Table 30–1). In some cases, it can be U, C, A or G *but in most cases it can be either the pair U and C or the pair A and G*. These relationships probably enable in some way the anticodon triplet in tRNA to recognize the codon triplet in mRNA. It has been found that each tRNA anticodon recognizes the several mRNA codons for one amino acid. Generally, one species of tRNA can recognize three codons which have the last base U, C and A as a group. *Codons ending in the base G are mostly specific for species of tRNA*. To explain the capacity of the third anticodon base to recognize several complementary bases in the codon (redundancy), Crick (1965) postulated the **wobble hypothesis**. According to this hypothesis, the first two bases of the anticodon are strictly standard for the first two constant bases of the codon and pair strongly with them. The third base of the codon is not so specific in its base pairing and may *wobble* (*pair loosely*) in pairing with the corresponding base in the anticodon. As a result of this, each tRNA recognizes several codons for its amino acid. Hence, the number of tRNAs needed for the translation of genetic code is considerably less than the number of codons. He discovered that if U is present at first position of anticodon, it can pair with either A or G at the third position of codon. Similarly, when C or A occurs in the 5′ position of the anticodon, it can pair only with G or U, respectively, in the 3′ position of a codon. Transfer RNAs containing either G or U in the 5′ (or wobble) position of the anticodon can each pair with 2 different codons, whereas an inosine, (the deaminated form of adenine), in this position produces a tRNA that can pair with 3 codons differing in the 3′ base. The pairing relationships between first base of anticodon and third base of codon are given in Table 30–5.

Table 30–5. Wobble base-pairing hypothesis

*Anticodon base** *(first base)*	*Codon base (third base)*
U	A, G
C	G only
A	U only
G	U, C
I (Inosine, resembles G)	U, C, A

* The first base of anticodon pairs with the third base of codon.

Fig. 30–9 presents examples of a standard base pair and two wobble base pairs.

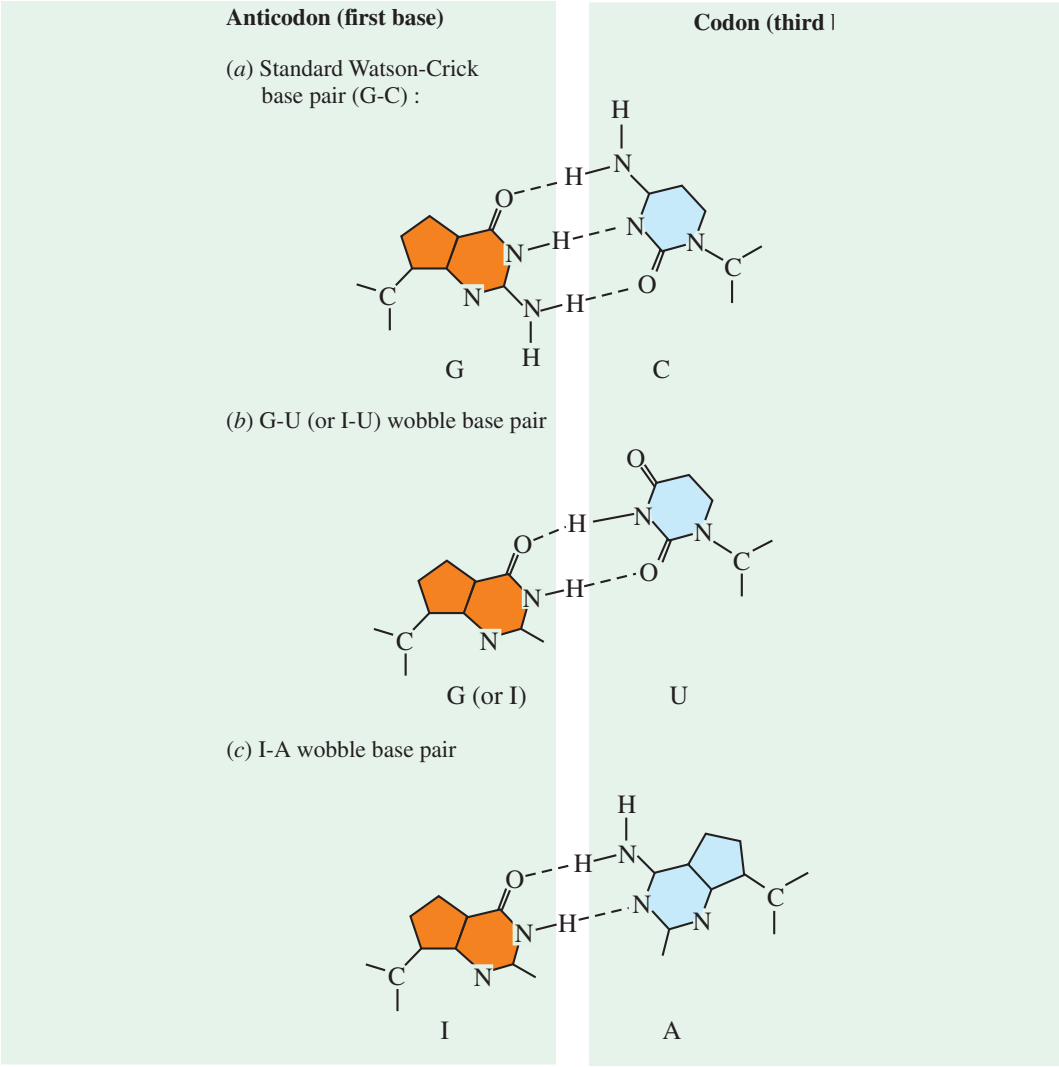


Fig. 30–9. Examples of standard and wobble base pairs

The standard (*a*) and wobble (*b* and *c*) base pairs are formed between the first base in the anticodon and the third base in the codon.

This kind of wobbling allows economy of the number of tRNA molecules, since several codons meant for same amino acid are recognized by the same tRNA. For instance, (i) anticodon IGC can recognize codons GCU, GCC and GCA (coding for alanine), (*ii*) IGA can recognize UCU, UCC and UCA (coding for serine), and (*iii*) IAC can recognize GUU, GUC and GUA (coding for valine). *Block explained multiple codon recognition on the basis of tautomeric shift model*. In some cases, the third base of the anticodon is an unusual base instead of one of the four common bases (Table 30–6).

Table 30–6. Codons and anticodons for some amino acids

Amino acid	*Alanine*	*Serine*	*Isoleucine*	*Phenylalanine*	*Tyrosine*	*Valine*
Codon	U	U	U	U	U	U
(5′–3′)	GCC	UCC	AUU	UUU	UAC	GUU
	A	A	A			A
	G	G				G
Anticodon (3′–5′)	CGI	AGI	UAI	AAMG	APSUG	CAI

PREFERENTIAL CODON USAGE

The fact that many of the amino acids have more than one codon assignment presents the question : Are each of the alternative codons within a set used with equal frequency, or are some of the codons in a set used more frequently than others ? From the study of gene sequences, it is now apparent that some codons are used more frequently than others. Surprisingly, in different species, the preferred codon usage is not the same for the same amino acids (Table 30–7). For example, even though leucine (Leu) has 6 anticodon assignments, *E. coli* uses CUG over 85% of the time, whereas yeast uses UUG over 85% of the time. As expected, preferential codon usage is correlated with the abundance of the respective tRNA species. The more frequently a codon is used, the more that tRNA species is present in the cell.

Additionally, within an organism, some codons are used more frequently for certain types of genes. For example, genes that are expressed at high levels preferentially use one codon in a set (*e.g.*, only one of the 6 codon assignments for leucine). Genes for less abundant proteins use a larger set of codons and show less preference toward the set encoding the more abundantly synthesized proteins. The reason for this difference in codon usage is not clearly known.

Table 30–7. Comparison of preferential codon usage for some amino acids

Codon for	*E. coli*	*Yeast*	*Euglena chloroplast*
Arginine (6)*	CGC	AGA	CGU/CGC/AGA
Leucine (6)	CUG	UUG	UUA/UUG/CUU
Serine (6)	UCU/UCC/AGC	UCU/UCC	UCU/UCA/AGU
Proline (4)	CCG	CCA	CCU/CCA
Tyrosine (4)	UAC	UAC	UAU/UAC
Lysine (2)	AAA	AAG	AAA

*The number in parenthesis represents the number of codons for that amino acid.

Preferred codons are those used more than 85% of the time. For example, UCU, UCC and AGC are used more than 85% of the time for serine in *E. coli.*

Another observation relating to codon usage is that codons with the forms NCG and NUA (n = any nucleotide) are avoided. It is speculated that it may be related to the frequent methylation of cytosine in CG dinucleotides and the tendency for methyl-CG to mutate by deamination to TG. However, this reasoning would apply only to eukaryotes because DNA methylation is minimal in prokaryotes.

Preferential codon usage may have some practical applications when it comes to transferring genes between different organisms and species. For example, the insulin gene from humans has been put into fast-growing *E. coli* to produce human insulin cheaply and in bulk.

MUTATIONS AND GENETIC CODE

Most of the work earlier discussed in this chapter deals with the study of genetic code in cell-free systems. Therefore, it could be questioned whether or not this information would apply to the living systems. With the study of certain mutants, it was possible to show that *the genetic code deciphered using cell-free systems applies to the living systems also.*

There are two kinds of mutations which have played a very significant role in the study of the genetic code. These are (*i*) frame shift mutations and (*ii*) base substitutions.

A. Frameshift Mutations

The genetic message, once initiated at a fixed point, is read in a definite frame in a series of three-

letter words. The framework would be disturbed as soon as there is a deletion or addition of one or more bases. When such frame shift mutations were intercrossed, in certain combinations, they gave wild type. It was concluded that one of them was deletion and the other an addition, so that the disturbed order of the frame due to one mutation will be restored by the other (Fig. 30-10).

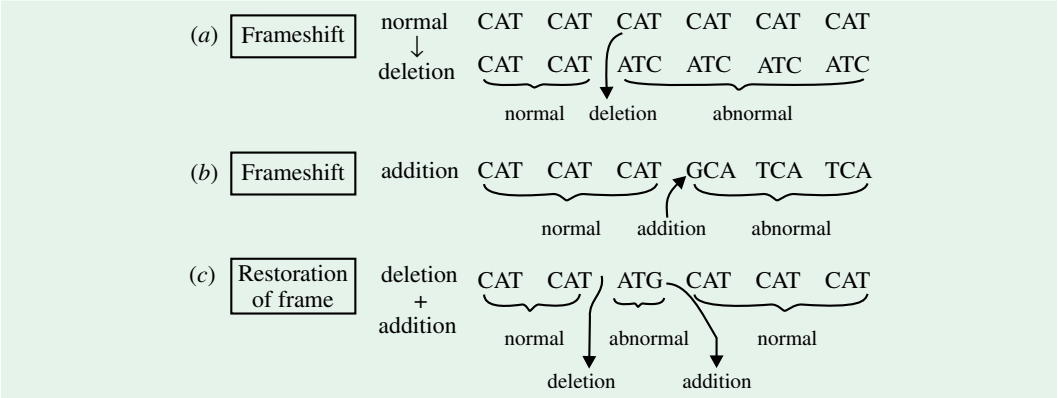


Fig. 30–10. Frame-shift mutations due to deletion and addition

B. Base Substitutions or Amino Acid Replacements

If in a mRNA at a particular point, one base is replaced by another without any deletion or addition, the meaning of one codon containing this altered base will ordinarily change and in place of a particular amino acid at a particular position in a polypeptide, another amino acid will be incorporated. Such mutations have been studied in detail in a protein enzyme known as *tryptophan synthetase*. By working out the altered amino acid sequences in the polypeptides, some conclusions regarding the possible changes in the base-sequence in RNA can be easily made if the dictionary of genetic code is available. *This has been done in tryptophan synthetase and also in case of hemoglobin, using sickle cell anemia.* From the analysis of many different species, including bacteria, protozoa, fungi, plants and animals, researchers have found that the GC is nearly universal only a few rare exceptions to the GC have been noted. These are described in Table 30–8.

Table 30–8. Exceptions to the genetic code

Codon	*Usual Use*	*Alternate Use*	*Where Alternate Use Occurs*
AGA AGG	Arg	Stop, Ser	Some animal mitochondria, some ptorozoans
AUA	Ile	Met	Mitochondria
CGG	Arg	Trp	Plant mitochondria
CUU CUC CUA CUG	Leu	Thr	Yeast mitochondria
AUU CUG UUG	Ile Val Leu	Start(N-f Met)	Some prokaryotes[a]
UAA UAG	Stop	Glu	Some protozoans
UGA	Stop	Trp Selenocysteine	Mitochondria, mycoplasmas *E.coli**

* Depends on context of message, other factors.

NEW GENETIC CODES

A. Genetic Code in Mitochondria

Earlier we described that the genetic code is universal. It was also believed that the genetic code does not undergo any kind of evolution and, therefore, should be static. However, during the last few years, it was shown that variations in genetic code are found in mitochondria, particularly studied in yeast and mammals. In the case of yeast mitochondria (Fig. 30-11), UGA codes for tryptophan, although in the nuclear genes, UGA is a termination codon. Similarly, the codons beginning with CU represent Thr instead of Leu and the AUA codon represents Met instead of Ile. It has also been shown that although a new genetic code may exist in mitochondria, but mitochondria in all organisms may not have the same genetic code. For instance, UGA in mitochondria does not always code for tryptophan. In the gene coding for maize cytochrome oxidase subunit II, UGA codon is not present and tryptophan is coded by codon CGG (coding for arginine in the nuclear genes) codes for methionine in mitochondria of mammals, *Xenopus*, yeast and *Drosophila,* but not in *Neurospora* and *Aspergillus*.

First Position (5′ end)	Second Position: U	Second Position: C	Second Position: A	Second Position: G	Third Position (3′ end)
U	UUU, UUC: Phe AAG UUA, UUG: Leu AAU*	UCU, UCC, UCA, UCG: Ser AGU	UAU, UAC: Tyr AUG UAA, UAG: STOP	UGU, UGC: Cys ACG UGA: Trp ACU* UGG	U, C, A, G
C	CUU, CUC, CUA, CUG: Thr GAU	CCU, CCC, CCA, CCG: Pro GGU	CAU, CAC: His GUG CAA, CA G: Gln GUU*	CGU, CGC, CGA, CGG: Arg GCA[b]	U, C, A, G
A	AUU, AUC: Ile UAG AUA, AUG: Met UAU[a]	ACU, ACC, ACA, ACG: Thr UGU	AAU, AAC: Asn UUG AAA, AAG: Lys UUU*	AGU, AGC: Ser UCG AGA, AGG: Arg UCU*	U, C, A, G
G	GUU, GUC, GUA, GUG: Val CAU	GCU, GCC, GCA, GCG: Ala CGU	GAU, GAC: Asp CUG GAA, GA G: Glu CUU*	GGU, GGC, GGA, GGG: Gly CCU	U, C, A, G

Fig. 30-11. The genetic code of yeast mitochondria

The codons (5′ ⟶ 3′) are at the left and the anticodons (3′ ⟶ 5′) are at the right in each box.
* designates U in the 5′ position of the anticodon that carries the —$CH_2NH_2CH_2COOH$ grouping on the 5′ position of the pyrimidine.

[a] Two tRNAs for methionine have been found. One is used in initiation and one is used for internal methionines.
[b] Although an Arg tRNA has been foundin yeast mitochondria, the extent to which the CGN codons are used is not clear.

(*Adapted from S G Bonitz et al, 1980*)

It has also been shown that in mitochondria, a number of single tRNA species with U in the wobble position (first base of anticodon pairing with the third base in codon) read all four codons in a family. It is also shown that there are only 22 tRNAs in mitochondria as against 55 tRNAs in universal code, and these are adequate for reading 60 codons, the remaining four being termination codons.

B. Genetic Code in Ciliate Protozoa

In 1986, a different genetic code was shown to be present in ciliate protozoa (*Mycoplasma capricolum*). In this genetic code, codons UAA and UAG specify glutamine instead of stop signals. In future, more such cases may be discovered, showing diversity in the genetic code.

OVERLAPPING GENES

Earlier, we stated that the genetic code is nonoverlapping– each ribonucleotide in an mRNA is part of only one triplet. However, this characteristic of the code does not rule out the possibility that a single mRNA may have multiple initiation points for translation. If so, these points could theoretically create several different reading frames within the same mRNA, thus specifying more than one polypeptide. This concept of overlapping genes is illustrated in Fig. 30–12(*a*).

That this might actually occur in some viruses was suspected when phage $\phi \times 174$ was carefully investigated . The circular DNA chromosome consists of 5386 nucleotides, which should encode a maximum of 1795 amino acids, sufficient for 5 or 6 proteins. However, this small virus in fact synthesizes 11 proteins consisting of more than 2300 amino acids. A comparison of the nucleotide sequence of the DNA and the amino acid sequences of the polypeptides synthesized has clarified the appearent paradox. At least four cases of multiple initiation have been discovered, creating overlapping genes (Fig. 30–12(*b*)).

The sequences specifying the K and B polypeptides are initiated with separate reading frames within the sequence specifying the A polypeptide. The K sequence overlaps into the adjacent sequence specifying the C polypeptide. The E sequence is out of frame with, but initiated in that of the D polypeptide. Finally, the A′ sequence, while in frame, begins in the middle of the A sequence. They both terminate at the identical point. In all, seven different polypeptides are created from a DNA sequence that might otherwise have specified only three (A, C and D).

A similar situation has been observed in other viruses, including phage G4 and the animal virus SV 40. Like $\phi \times 174$, phage G4 contains a circular single-stranded DNA molecule. The use of overlapping reading frames optimizes the use of a limited amount of DNA present in these small viruses. However, such an approach to storing information has a distinct disadvantage in that a single mutation may affect more than one protein and thus increase the chances that the change will be deleterious or lethal. In the case we just discussed, a single mutation at the junction of genes A and C could affect three proteins (the A, C, and K proteins). It may be for this reason that overlapping genes are not common in other organisms.

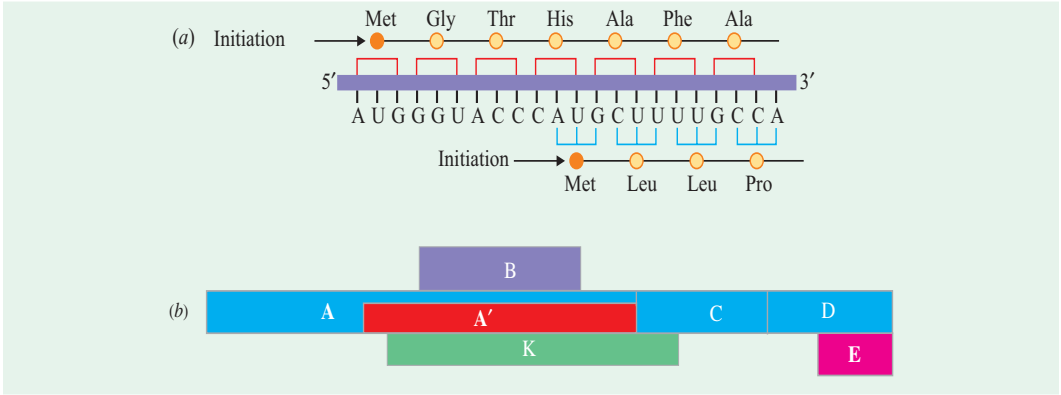


Fig. 30–12. An illustration of the concept of overlapping genes

(*a*) An mRNA sequence initiated at 2 different AUG positions out of frame with one another will give rise to 2 distinct amino acid sequences.

(*b*) The relative positions of the sequences encoding 7 polypeptides of the phage $\phi \times 174$.

EVOLUTION OF THE GENETIC CODE

Two types of theories have been putforth to explain the evolution of the genetic code.

Mechanistic models

These models suggest that the original code was based upon some physico-chemical relationship between the amino acid and its codon. **Woese** (1967-68) had proposed that the original code might have been determined by a stereochemical fit between an amino acid and its codon, the nucleic acid acting as a physical template for assembly of the amino acids. Initially, the code might commence in an autocatalytic cycle in which a polynucleotide and a polypeptide assisted each other to replicate, so that their relationship was reciprocal, only subsequently becoming unidirectional. Nucleotides and amino acids involved under primitive conditions were not the same as those employed today and this creates difficulty in confirming the primitive code. **Lacey and Pruit** (1969) have demonstrated the formation of a complex between poly-L-lysine and mononucleotides. It is a controversial matter whether an amino acid could interact with its codon with sufficient stereochemical specificity to fit to account for the development of the code. Nevertheless, it is an attractive hypothesis.

Stochastic models

These models do not depend on physico-chemical relationship. It is presumed that the initial set of codon assignments was free to vary. Selective forces would modify it in the course of evolution and continued unchanged in all organisms descended from a single inbreeding population. Pure mechanistic and pure stochastic models differ in the nature of selective forces responsible for the evolution of the code.

Sonneborn (1965) proposed a mutational-buffer theory. According to him, mutations resulting from substitution of a single nucleotide by another tend to exert a deleterious effect upon the cell and if the cells can reduce the burden of harmful mutations, they are likely to be at a selective advantage. This theory is purely stochastic as it demands that the codons for similar amino acids should be related to each other. This model suffers from following two drawbacks :

1. It does not account for the universality of the code.
2. The model requires codons to change their meanings during the course of evolution.

Woese (1965, 1967, 1968) proposed an alternative model which suggests that the selective forces in evolution operate to minimise errors made in the translation of codons into amino acids. *This is not a purely stochastic hypothesis as it depends to some extent upon mechanistic features involved in translation.* Woese argues that translation is highly evolved today but primitive cells must have had different systems. No experimental evidence is, however, possible to confirm the primitive translation machinery. Woese further suggests that evolution would favour a codon dictionary in which more error-prone codons are represented by amino acids which are functionally less important. He observed that the probability of misreading occurring at each of the three nucleotides in a codon are 100 : 10 : 1 for positions III : I : II respectively. Thus, the third position, being most error-prone, is the most degenerate so that codons differing in the base at this position tend to be assigned to related amino acid. The second base is the least error-prone and is probably least involved in such restraints.

Crick (1968) has put forward a model in which the code might start in primitive form, a small number of triplets coding for only a few amino acids and assignment of codons to amino acids would become nonsense and probably lethal, on switching to the triplet code. Nobody knows which amino acids were present in the initial code. It seems probable that the present complex species might not have existed under prebiotic conditions. Through a series of changes, the nature of which is difficult to fathom, the final code would evolve as new amino acids replaced some of the primitive ones. This is possible if the new amino acids are related and the organism codes for a few rather crude and simple proteins. As the number and complexity of proteins increased, a situation might arise when substitution of new amino acids may prove lethal and at this point the code would be 'frozen'. This

theory is termed a '*frozen accident model*'. But in order to explain universality, it is again necessary to postulate that this must have happened at an early stage of evolution where all cells descended from one population were involved.

REFERENCES

1. **Barrell BG, Banker AT, Drouin J :** A different genetic code in human mitochondria. *Nature* **282 :** *189-194, 1979.*
2. **Barrell BG, Air G, Hutchinson C :** Overlapping genes in bacteriophage $\phi \times 174$. *Nature* **264 :** *34 - 40, 1976.*
3. **Chambron P :** Split genes. *Scientific American. 244 (May) : 60-71, 1981.*
4. **Crick FHC Barnett L, Brenner S, Watts-Tobin RJ :** General nature of the genetic code for proteins. *Nature, 192 : 1227-1232., 1961.*
5. **Crick FHC :** Codon–anticodon pairing : The wobble hypothesis. *J.Mol. Biol.* **19 :** *548 - 55, 1966 b.*
6. **Crick FHC :** The genetic code. *Sci. America (Oct)* **207 :** *66-77, 1962.*
7. **Crick FHC :** The genetic code III. *Sci. Amer. (Oct.)* **215 :** *55-63, 1966a.*
8. **Gamow G :** Possible relation between DNA and protein structures. *Nature* **173 :** *318, 1954.*
9. **Jukes TH :** The genetic code. *Am. Sci.* **51 :** *227-245, 1963.*
10. **Khorana HG :** Polynucleotide synthesis and the genetic code. *Harvey Lectures.* **62 :** *79 - 105, 1967.*
11. **Khorana HG et, al :** Polynucleotide synthesis and the genetic code. *Cold Spring Harb. Sym. 31 : 39-49, 1967.*
12. **Knight RD, Freeland SJ, Landweber LF :** Selection, history and chemistry : the three faces of the genetic code. *Trends Biochem. Sci.* **24(6) :** *241-247, 1999.*
13. **Lewin B :** Genes VI. *6th ed. Oxford University Press, New York. 1997.*
14. **Nirenberg MW :** The genetic code : II *Scientific American.* (*March issue*) **190 :** *80-94, 1963.*
15. **Nirenberg MW, Leder P :** RNA code words and protein synthesis. *Science* **145 :** *1399 - 1407, 1964.*
16. **Ochoa S :** Synthetic polynucleotides and the genetic code (Symposium on Genetic Mechanics). *Fed. Proc. 22: 62-74, 1963.*
17. **Singer M, Berg P :** Genes and Genomes. *University Science Books, Mill Valley, CA . 1991.*
18. **Smithe MA et, al :** Direction of reading of the genetic message, II *Proc. Nat, Acad. Sci. 55 : 141-147, 1966.*
19. **Szathmary, E :** What is the optimum size for the genetic alphabet. *Proc. Natl. Acad. Sci. 89 : 2614-2618, 1992.*
20. **Woese CR :** The Genetic Code. *Harper & Row, New York. 1967.*
21. **Wong JT :** Evolution of the genetic code. *Microbiol. Sci.* **5(6) :** *174-181, 1988.*
22. **Ycas M :** The Biological Code. *North Holland, Amsterdam. 1969.*

PROBLEMS

1. (*a*) Write the sequence of the mRNA molecule synthesized from a DNA template strand having the sequence

5′-ATCGTACCGTTA-3′

(*b*) What amino acid sequence is encoded by the following base sequence of an mRNA molecule ? Assume that the reading frame starts at the 5′ end.

5′-UUGCCUAGUGAUUGGAUG-3′

(*c*) What is the sequence of the polypeptide formed on addition of poly (UUAC) to a cell-free protein-synthesizing system ?

2. The code word GGG cannot be deciphered in the same way as can UUU, CCC, and AAA, because poly(G) does not act as a template. Poly(G) forms a triple-stranded helical structure. Why is it an ineffective template ?

3. Synthetic RNA molecules of defined sequence were instrumental in deciphering the genetic code. Their synthesis first required the synthesis of DNA molecules to serve as a template. Har Gobind Khorana synthesized, by organic-chemical methods, two complementary deoxyribonucleotides, each with nine residues : $d(TAC)_3$ and $d(GTA)_3$. Partly overlapping duplexes that formed on mixing these oligonucleotides then served as templates for the synthesis by DNA polymerase of long, repeating double-helical DNA chains. The next step was to obtain long polyribonucleotide chains with a sequence complementary to only one of the two DNA strands. How did he obtain only poly(UAC) ? Only poly(GUA) ?

4. In a nonoverlapping triplet code, each group of three bases in a sequence ABCDEF... specifies only one amino acid—ABC specifies the first, DEF the second, and so forth– whereas, in a completely overlapping triplet code, ABC specifies the first amino acid, BCD the second, CDE the third, and so forth. Assume that you can mutate an individual nucleotide of a codon and detect the mutation in the amino acid sequence. Design an experiment that would establish whether the genetic code is overlapping or nonoverlapping.

5. Proteins generally have low contents of Met and Trp, intermediate ones of His and Cys, and high ones of Leu and Ser. What is the relation between the number of codons of an amino acid and its frequency of occurrence in proteins. What might be the selective advantage of this relation.

6. Crick, Barnett, Brenner, and Watts-Tobin, in their studies of frameshift mutations, found that either three pluses or three minuses restored the correct reading frame. If the code were a sextuplet (consisting of six nucleotides), would the reading frame be restored by either of the preceding combinations ?

7. When repeating copolymers are used to form synthetic mRNAs, dinucleotides produce a single type of polypeptide that contains only two different amino acids. On the other hand, using a trinucleotide sequence produces three different polypeptides, each consisting of only a single amino acid. Why ? What will be produced when a repeating tetranucleotide is used ?

8. In studies using repeating copolymers, ACA ... incorporates threonine and histidine, and CAACAA ... incorporates glutamine, asparagine, and threonine. What triplet code can definitely be assigned to threonine ?

9. In the triplet-binding technique, radioactivity remains on the filter when the amino acid corresponding to the triplet is labeled. Explain the basis of this technique.

10. (*a*) Shown here is a theoretical viral mRNA sequence :

5′–AUGCAUACCUAUGUGACCCUUGGA–3′

Assuming that it could arise from overlapping genes, how many different polypeptide, sequences can be produced ? Using Figure 30–2, what are the sequences ?

(*b*) A base substitution mutation that altered the sequence in (*a*) eliminated the synthesis of all but one polypeptide.

The altered sequence is shown here :

5′–AUGCAUACCUAUGUGACCCUUGGA–3′

Using Fig. 30-2, determine why ?

11. Define the process of transcription. Where does this process fit into the central dogma of molecular genetics ?

12. Describe the structure of RNA polymerase in bacteria. What is the core enzyme ? What is the role of the sigma subunit.

13. In a mixed copolymer experiment, messengers were created with either 4/5C: 1/5A or 4/5A: 1/5C. These messages yielded proteins with the following amino acid compositions.

4/5C : 1/5A		4/5A : 1/5C	
Proline	63.0 percent	Proline	3.5 percent
Histidine	3.0 Percent	Histidine	3.0 percent
Threonine	16.0 percent	Threonine	16.6 percent
Glutamine	3.0 percent	Glutamine	13.0 percent
Asparagine	3.0 percent	Asparagine	13.0 percent
Lysine	0.5 percent	Lysine	50.0 percent
	98.5 percent		99.1 percent

Using these data, predict the most specific coding composition for each amino acid.

14. What would be the effect on reading frame and gene function under the following conditions?

(*a*) Three bases were inserted together in the middle of the gene.

(*b*) Three bases were deleted from the gene.

(*c*) One base was inserted and another one deleted five bases downstream of the insertion.

15. You know the amino acid sequence of a protein coded for by a gene in *E.coli* and the very end of the sequence is

Pro–Trp–Ser–Glu

You find a mutant of this gene and the preceding sequence of the protein has changed to

Pro–Gly–Val–Lys–Met–Arg–Val

Explain what has happened. What is your prediction as to the effect on protein function?

16. How does aminoacyl-tRNA synthetase^{Leu} recognize only the tRNA^{Leu} family (of which there are six) to specifically attach leucine to the proper tRNA ?

17. Frameshift mutations frequently cause the translated protein to terminate downstream of the mutation. How do you explain this phenomenon ?