EVIDENCE FOR THE GENETIC CODE

The experimental evidence for the nature and meaning of the genetic code has been obtained through the following five different experimental approaches.

1. genetic analysis of deletion mutants of a cistron.
2. Incorporation of amino acids by synthetic polyribonucleotides.
3. Binding of known triplets to aminoacyl-tRNA.
4. Amino acid sequence generated by mRNA transcribed from synthetic DNA of known nucleotide sequence.
5. Comparison of nucleotide sequence of a cistron and the amino acid sequence of the polypeptide produced by it.

Genetic analysis of deletion and insertion mutants, spontaneous as well as induced by acridine dyes, of the *B* cistron of the *rII* locus of phage T_4 was reported by Crick *et al.* in 1961. Deletion or insertion of a single nucleotide in a polynucleotide sequence leads to a shift in the reading frame (***frame shift mutation***) after the point of deletion or insertion (Fig. 34.2). As a result, the reading of subsequent codons is out of phase, *i.e.*, the base sequence of each codon is now changed. The amino acid sequence of the resultant protein is, therefore, grossly changed and the protein is functionally inert. All the mutations tested in this study produced mutant *rII* phenotype. One of the mutations, FC0, was arbitrarily designated as +. Another mutation FC1, located at a different site, suppressed the effect of FC0 and produced wild type phenotype when both the mutations were present in the same chromosome (***cis configuration***), and was labelled as –. However, mutation FC40 could not suppress the effect of FC0, but suppressed the effect of FC1, and was hence classified as +. In this manner all the mutations were classified as + and –. In general, one + mutation located after a – mutation, or *vice versa*, was able to suppress the mutant effect and produce the wild type phenotype, two + or two – mutations present together still produced the mutant phenotype, *i.e.*, there was no suppression. But most importantly, *three + or three – mutations present in the* cis *position produced the wild type, and the mutant phenotype was suppressed.* ***Thus the codon is most likely a triplet*** (or, perhaps, it has a multiple of 3 bases).

The suppression of the mutant action of a + mutation by a – mutation located at a subsequent site depends on a commaless nature of the code. This is because deletion or insertion of a base would change the number of bases between two commas adjacent to the point of insertion/deletion (*e.g.*, XTAGX + T → XTTAGX, where X is the nucleotide representing the comma). If such were the case, reading of such a codon would not be possible and the codon would most likely act as a nonsense codon terminating the polypeptide chain, which obviously is not the case.

It the code were not degenerate, *i.e.*, each amino acid was coded by a single codon, frame-shift due to a deletion/insertion would produce many new codons in addition to the original 20 or so codons present. The new codons would obviously be nonsense codons, but this is not the case in most of such mutations (although some such mutations do lead to the formation of nonsense codons). Therefore, *most likely the code is degenerate.* It may be noted that each + mutation may represent a deletion (or insertion) of one or two bases in each case; the – mutations would then represent insertion (or deletion), *i.e.*, the opposite of the + mutations.

Synthetic polyribonucleotides were used as mRNA by Nirenberg and Matthei to determine the meaning of codons. Synthetic RNA is produced by the enzyme polynucleotide phosphorylase (discovered by Ochoa and his group in 1955), which does not use DNA as template and incorporates ribonucleotides in a random order. The first synthetic RNA used for protein synthesis in a cell-free system (containing ribosomes, tRNA's and other factors for portein synthesis) was poly-U; it directed the synthesis of polyphenylalanine, *i.e.*, a polypeptide containing only phenylalanine. This showed that, most likely, the triplet UUU codes for phenylalanine. Subsequently, AAA was shown to code for lysine, and CCC to specify proline, by employing poly-A

and poly-C, respectively, as mRNA. Poly-G forms multistranded helices due to hydrogen-bonding between guanine residues of different strands, hence it is unable to function as mRNA.

Subsequently, Nirenberg and coworkers and Ochoa and his group used polyribonucleotides containing mixtures of two, three or all the four nucleotides as mRNA. This yielded further information on the possible codons for certain amino acides. For example, poly-CG directs the incorporation of arginine and alanine, in addition to the proline coded by poly-C. Eight different codons are possible in poly-CG, *viz.*, CCC, CCG, CGC, GCC, CGG, GCG, GGC and GGG; the amino acids arg and ala are coded by one or more of these codons, except CCC. Some additional information on the assignment of codons to individual amino acids was obtained by varying the ratios of the different nucleotides in the mixed polyribonucleotides, but the code assignment was at best ambiguous. *Since all the four bases were not required for the incorporation of any of the amino acids, it became clear that the codon was either a doublet (containing two bases) or a triplet.*

A triribonucleotide cannot direct the synthesis of a polypeptide, but it leads to the binding of specific aminoacyl-tRNA to the ribosomes. This property is exploited by the ***Nirenberg-Leder technique*** for assigning specific codons to specific amino acids. Conventionally, a trinucleotide with 5′ terminal phosphate (*e.g.*, of U) is written as pUpUpU, while that with 3′-terminal phosphate is symbolised as UpUpUp. Trinucleotides with 5′-terminal phosphate are more active than those lacking this phosphate in inducing the binding of aminoacyl-tRNA to ribosomes, while those with 3-terminal phosphate and diribonucleotides are inactive. *This clearly showed that the codon had three bases, or was a triplet.*

In the Nirenberg-Leder technique, indvidual trinucleotides with 5′-terminal phosphate are added to a cell-free system supporting protein synthesis and containing a mixture of the 20 amino acids bound to their respective tRNAs. One of the 20 amino acids in the mixture is radioactive (^{14}C); a different radioactive amino acid is used in the 20 separate experiments set up using the same trinucleotide. The cell-free mixture is then filtered through a nitrocellulose filter which retains the ribosomes but allows the aminoacyl-tRNA to pass through. The ribosomes are then tested for the presence of radioactivity to detect the binding of the radioactive aminoacyl-tRNA, present in the mixture, to the ribosomes. These studies confirmed that UUU, CCC and AAA induced the binding of phenylalanine, proline and lysine, respectively, to ribosomes. Using this technique, most of the codons were assigned to specific amino acids by Nirenberg and coworkers, and by Khorana and his group.

The fourth approach is known as the ***Khorana technique.*** In this technique, long specific DNA sequences are first synthesized, which are then transcribed by DNA-directed RNA polymerase to yield complementary RNA molecules. These RNA molecules are used as mRNA to synthesize polypeptides, whose amino acid sequence is then determined. The nucleotide sequence is then compared with the amino acid sequence of the polypeptide it has produced to deduce the specific codons for different amino acids. *This technique was exploited by Khorana and associates to confirm that the code is commaless, nonoverlapping and degenerate, and that each codon*

is a triplet. It also confirmed the codon assignment through the Nirenberg-Leder technique and resolved the disputes regarding assignment of certain codons.

The ***in vivo evidence*** for the genetic code is both indirect and direct. The indirect evidence is based on the analysis of amino acid replacements in a protein produced by different mutations/reversions. The codons for the replaced amino acid and for the amino acid replacing it are then compared to see if a change in a single base would produce the observed change. In all such cases analysed, the observed amino acid replacements can be easily accounted for by the proposed genetic code (Fig. 34.1). The direct *in vivo* evidence for the genetic code comes from a comparison of the base sequence of a cistron and the amino acid sequence of the polypeptide produced by it. Such a study has been done with the small RNA bacteriophage MS2 which has only three cistrons. One of these cistrons produces the coat protein which has 129 amino acids. The cistron, as expected, has 387 nucleotides. There is a complete agreement between the amino acid sequence expected from the base sequence of this cistron and the amino acid sequence of the head protein actually produced by it. In other words, each codon of the cistron codes for precisely the same amino acid as expected from the coding dictionary listed in Fig. 34.1. This analysis also confirmed beyond any doubt the triplet, nonoverlapping, commaless nature of the genetic code. The codon AUG evidently functions as the polypeptide initiation codon, while chain termination is due to two successive nonsense codons (UAA and UAG). Further, all the 64 codons are not used by MS2; some codons for an amino acid are used to the exclusion of the others for the same amino acid. For example, isoleucine is coded by AUU and AUG in MS2, and the codon AUA is not used. Thus degeneracy of the genetic code in an organism appears to be somewhat restricted.