THE GENETIC CODE

by

M. W. Nirenberg

National Heart Institute

National Institutes of Health

Bethesda, Maryland

DRAFT OF THE MANUSCRIPT 9/15/67

I. INTRODUCTION

In recent years studies on the genetic code, protein synthesis, and regulation of protein synthesis have expanded to such proportions that investigators in other fields and graduate students often find it difficult to assimilate unrelated facts and to formulate general principals from them. In writing this chapter the objectives have been to first, the approaches which were used to decipher the code, second, the nature of the code and the evidence for it, and third, the apparent logic of the code's design. An attempt is made to formulate general principals and the logical design of the code.

This chapter is meant to be a discussion of the code, how it was translated and the experimental data. In writing the chapter an attempt has been made to formulate general principals from the data and to state the apparent logic which underlies the design of the code. Coverage of topics has been selective rather than comprehensive, and in some ways, the chapter is an essay than a review.

The data which is now available on the code, the structure and function of nucleic acids and protein, and the process of protein synthesis

is so extensive that it was deemed essential from the outset to be selectively rather than comprehensive in discussing the data.

An attempt has been made in this chapter to survey the genetic code, to concentrate on fundamental principals, especially the apparent design of the code and the logic which is employed. Coverage is selective rather than comprehensive.

B. Base Compositions of Codons

Synthetic RNA preparations containing all possible base combinations have been used as templates for ~~protein~~ polypeptide synthesis in vitro. In practice, the major factor ~~which limits~~ limiting the sensitivity of the assay is the ~~presence~~ amount of endogenous mRNA in E. coli extracts. As shown in Fig. _____, the level of endogenous mRNA is greatly reduced by ~~incubating~~ incubation of E. coli extracts in the presence of DNase and all components required for protein synthesis until amino acid incorporation ceases. ~~Extracts then can be dialized and stored until needed~~. Protein synthesis then is almost completely dependent upon the addition of mRNA.

Optimal conditions for in vitro protein synthesis stimulated by synthetic mRNA were determined ( ), and methods were devised for rapidly washing radioactive protein precipitates on cellulose nitrate filters ( ). Most radioactive ~~protein~~ polypeptide products are washed with 10% trichloroacetic acid. Those rich in proline are washed with 20% trichloroacetic acid ( ), whereas lysine-rich ~~proteins~~ products are washed with a solution containing sodium tungstate and trichloroacetic acid ( ).

The specificity of randomly-ordered RNA templates for amino

acid incorporation into ~~protein~~ polypeptide chains has been studied extensively with extracts of E. coli ( ). A summary of the minimum ~~kinds of bases which are required to code for each amino acid~~ base compositions of randomly-ordered polynucleotides that correspond to the amino acids is shown in Table 1. A polynucleotide ~~with~~ containing one kind of base ~~usually is a~~ template for ~~one~~ each amino acid ~~polypeptide~~, a single amino acid. Little template activity ~~was~~ is detected with poly G; presumably G-G interactions inhibit the template activity of RNA (discussed in a later section).

A polynucleotide with two kinds of bases contains eight triplets; six triplets with two kinds of bases, and two triplets with one ~~kind of~~ base. For example poly UC contains:

UUC UUU
UCU CCC
CUU
CCU
CUC
UCC

Three preparations, poly UC, poly CG, and poly AG, ~~are active~~ each serves as templates for four amino acids per polypeptide. ~~three~~ Each one of other polynucleotide preparations (poly UA, poly UG, and poly CA) ~~serves as~~ is a template for six amino acids.

~~There are~~ The four possible polynucleotides ~~with~~ containing three kinds of bases are poly UAG, poly UCG, poly UCA, and poly CAG. A polynucleotide with three

kinds of bases, ~~for example,~~ such as poly UAG, resembles a mixture of seven polynucleotides as follows: poly U, poly A, poly G, poly UA, poly UG, poly AG, poly UAG. ~~Poly~~ Randomly polymerized UAG contains ~~three~~ twenty-seven triplets: with one kind of base, eighteen ~~triplets~~ with two kinds of bases, and six ~~triplets~~ with three kinds of bases. ~~a total of twenty-seven triplets.~~

Each three-base polynucleotide stimulated the incorporation into protein of at least ten ~~or more~~ amino acids. Four ~~three-base~~ codons were found which could not be accounted for by ~~one-base, or two-base codons;~~ codons with one or two kinds of bases; ~~poly~~ UAG ~~was a template for~~ appears to code for methionine and aspartic acid; and ~~poly~~ CAG was a template for serine and aspartic acid. ____additional three-base codons were assigned to amino acids. ~~Such~~ These results clearly demonstrate specificity of tRNA for amino acid as a function of the ~~template specificity of codons and~~ the minimum kinds of bases which are ~~may be~~ present in codons for each amino acid. ~~In addition the results show that the code is degenerate because amino acids such as leucine, arginine and serine respond to several polynucleotides which differ in base content.~~

With additional data it is possible to ~~derive~~ deduce the ~~relative number~~ ~~proportion of~~ bases ~~in many~~ per codon as well as the kinds of bases ~~which~~ that are present. Base compositions of RNA codons are derived as follows: The base composition of a polynucleotide ~~can be~~ is determined ~~quite accurately~~ experimentally; then the expected frequency of each doublet or triplet is calculated easily once the base ratio of a randomly-ordered polynucleotide is known. By synthesizing a series of polynucleotides, each containing identical bases, but ~~with~~ in different proportions ~~of bases~~, and subsequently determining the relative proportions of amino acids directed into ~~protein~~ polypeptide material by each polynucleotide, ~~one can estimate the~~ base compositions of ~~the~~ codon, as well as the number of nucleotides per codon ~~can be~~ could be estimated.

~~It is difficult to compare directly the~~ a direct comparison of template efficiency of ~~one~~ different polynucleotide preparations ~~with another, for~~ is difficult since the efficiency of each preparation may be influenced strikingly by factors other than base composition, such as, the conformation of the RNA in solution, the presence or absence of terminal phosphate, ~~its molecular weight~~, the number of base residues per molecule and so forth. These factors will be discussed in later sections. However, by determining the amount of each amino acid incorporated into protein due to the addition of a particular polynucleotide preparation, ~~can be determined~~

the relative proportions of amino acids incorporated under the influence of different polynucleotide preparations can be compared.

In Table ___ 3 shown an example of data obtained with a poly AC preparation; similar data were obtained with four other poly AC preparations, each differing in base ratio ( ). The four possible doublet permutations do not contain enough specific information to code for the six amino acids directed into protein by poly AC, whereas, the information content of the eight possible triplets is adequate. If every triplet were read, some amino acids would respond to two or more codons. In such cases the sum of the triplet frequencies would then be compared with the corresponding amino acid incorporation data. For example, if (ACA) and (ACC) both corresponded to the same amino acid, the sum of their frequencies would be 24.9 percent, which could not be distinguished from the frequency of the doublet, (CA), which is also 24.9 percent.

The relation between theoretical frequency and the experimentally determined frequency of amino acid incorporation into protein is shown in Fig. ____. The data demonstrate clearly that histidine, asparagine and glutamine are coded by triplets containing two C's and one A. The composition of a histidine codon is (CAC), an asparagine codon,

(AAC) and a glutamine codon, (CAA). Threonine responds either to two triplets, one of base composition (ACA) and the other (ACC), or to the doublet (AC).

As shown in Fig. ___, proline responds to two triplets, CCC and (CCA), or to the CC doublet, and ~~that~~ lysine responds to the triplet AAA. ~~In this way~~ every triplet base ~~composition~~ combination in poly AC was assigned to an amino acid as follows:

Proline	CCC (CCA)
Histidine	(CAC)
Threonine	(ACC) (ACA)
Glutamine	(CAA)
Asparagine	(AAC)
Lysine	AAA

In this way the nucleotide compositions of approximately 50 codons were estimated by Ochoa ( ), Nirenberg ( ) and their coworkers. A summary of the ~~is~~ is shown in Table ____. Tentative base compositions were estimated for many codons containing three different bases. Most amino acids were found to be coded by multiple words (Triplets). Since synonym codons often differ by only one base, the bases which were common to each synonym codon were assumed to occupy the same position within each triplet.

Similar results were obtained in both laboratories, although extracts were prepared from E. coli B in the Ochoa laboratory and from E. coli ~~and~~ W3100

(a K12 strain) in the NIH laboratory.

C. Nucleotide Sequence of Codons

~~1. The Effect of Trinucleotides upon AA-tRNA Binding to Ribosomes~~

Each of the 64 trinucleotides have been synthesized, and assayed ~~as a~~ for template capacity in the ~~for~~ binding of E. coli AA-tRNA. Since the initial studies showed that AA-tRNA for some amino acids binds to ribosomes in response to trinucleotides at 0.02-0.03 M Mg^{++}, but not at 0.01 M Mg^{++}, (Nirenberg and Leder, 1964; Leder and Nirenberg, PNAS) a relatively high Mg^{++} concentration (0.03 M) was selected for the initial survey of trinucleotide-ribosome-AA-tRNA complexing. All responses found at 0.03 M Mg^{++} then were reassessed at 0.01 and 0.02 M Mg^{++}.

Summaries of responses of unfractionated E. coli AA-tRNA are shown in Tables ___ and ___. (Continued on next page)

Most trinucleotides have been assayed for template specificity with 20 AA-tRNA purified preparations from E. coli, each acylated with one radioactive and 19 unlabeled amino acids ( ). In surveying trinucleotide specificity, unfractionated AA-tRNA ~~usually is~~ generally has been used only initially, because the various species of tRNA compete with one another during the formation of AA-tRNA-codon complexes and the specificity of codon recognition can be altered by changing the concentrations of two or more species of tRNA.

Insert discussion of tables + data

Almost all triplets ~~were found to~~ correspond to amino acid.

Synonym codons ~~were found to be~~ are logically related to one another, and in most cases, ~~synonym codons~~ differ only in the base occupying the third position of the triplet. Only four unique patterns of degeneracy were found, each pattern determined by the kinds of bases which occupy the third positions of synonym triplets. Patterns of alternate third bases are:

1) G
2) U = C
3) A = G
4) U = C = A

check Table for data C. as

A fifth pattern, U = C = A = G, ~~is found~~ occurs which is not necessarily unique, because this pattern would result if two simpler patterns were present, such as [(U = C) + (A = G)] or [(U = C = A) + (G)].

Codons specifying the initiation of protein synthesis may contain alternate bases at the first rather than the third position of the codons. For example, N-formyl-Met-tRNA responds to AUG, GUG, and possibly also UUG (discussed under Punctuation).

IK

One consequence of logical degeneracy is that many mutations leading to single base replacements in DNA at sites corresponding to third bases of mRNA codons may not result in amino acid replacement in protein. Hence, many mutations ~~thus~~ are "silent." The code appears to be arranged so that the effects of base replacements in DNA, or erroneous translation of a base in mRNA, often is minimized. Possible amino acid replacements in protein which would occur as a result of single base changes can be read in Table ___ by moving horizontally or vertically from the amino acid in question, but not diagonally. ~~The code appears to be arranged so that the effects of some errors may be minimized, since~~ amino acids which are structurally or metabolically related often correspond ~~to similar~~ RNA codons. For example, Asp-codons, GAU and GAC, are similar to Glu-codons, GAA and GAG; Ser-codons are related to Thr-codons, and so forth.

demonstrating base sequences of codons

These results confirm 46 of the 53 codon base compositions which had been assigned on the basis of studies with synthetic, randomly-ordered polynucleotides and the cell-free protein synthesizing systems of the

Matthaei and coworkers determined upon AA-tRNA binding to ribosomes the effects of sixteen polynucleotide preparations, each with a different 5'-terminal doublet followed by approximately one hundred C residues, such as AU(C)$_{100}$ ~~upon AA-tRNA binding to ribosomes were determined~~ (Matthaei). The results are shown in Table ___. Each RNA preparation markedly stimulated binding of Pro-tRNA and Ser-tRNA. Ala-tRNA also responded to most of the polynucleotides, ~~also~~. The ~~high~~ marked responses of Pro-tRNA to every polynucleotide demonstrates a high out-of phase response to CCC residues. The very high response of Ser-tRNA to polynucleotides ~~is twice~~ was unexpected. Each RNA preparation contains three triplets, depending upon the phase of codon recognition. For example, poly UUCCC(C)$_{100}$ contains the triplets UUC, UCC, and CCC, according to the reading phase and stimulates binding of Pro-tRNA, Ser-tRNA, Phe-tRNA and Ala-tRNA to ribosomes. In some cases, bases one to three are recognized preferentially; in other cases, bases two to four are preferred. Therefore, phasing preferences depend upon the triplet rather than the exact location of the triplet in the 5'-terminal region. ~~Although such data are of considerable interest in considering regard to the preferred modes of phasing codon recognition, interpretation of codon base sequence often is complicated by the phasing problem.~~

~~Nevertheless~~, responses of AA-tRNA to approximately half of the sixteen codons tested are sufficiently high that base sequence assignments can be derived readily. However, interpretation of the remaining data is complicated by the phasing problem. Responses of AA-tRNA to polynucleotides agree well with the responses of AA-tRNA to trinucleotides in ribosome binding if the assumption is made that almost every possible kind of triplet in each polynucleotide is recognized to a greater or lesser extent. However, three differences in response of AA-tRNA to polynucleotides and trinucleotides should be noted; Gly-tRNA does not respond to $GGC(C)_{100}$, but does respond to the trinucleotide, GGC; Leu-tRNA does not respond to $CUC(C)_{100}$ but does respond to randomly-ordered poly UC, although responses to the trinucleotide, CUC, may be difficult to detect with unfractionated Leu-tRNA ( ); and finally, high responses of Ser-tRNA to polynucleotides are not observed with trinucleotides in the ribosome binding system.

Khorana and his colleagues ( ) synthesized polynucleotides with repeating doublet, triplet, or tetramer sequences ~~were synthesized by Khorana and his colleagues ( )~~ and ~~were~~ used them to stimulate amino acid incorporation in E. coli extracts. ~~A summary of~~ the results is summarized in Table ___. ~~is shown in Table ___~~. RNA preparations which do not contain an initiator codon, such as AUG or GUG, are translated in almost every possible phase during protein synthesis in E. coli extracts. RNA with a repeating doublet

sequence contains two triplets in alternating sequence, and therefore is a template for ~~protein~~ polypeptide containing two amino acids in alternating sequence. Most RNA preparations with a repeating triplet sequence are read in three phases, each phase corresponding to a different triplet. For example, poly UUC stimulates the incorporation of radioactive phenylalanine, serine and leucine into protein. Therefore, poly UUC resembles a mixture of poly UUC, poly UCU, and poly CUU; if the reading phase were $(-UUC \cdot UUC-)_n$, the protein product would be polyphenylalanine; if the phase of reading were $(-UCU \cdot UCU-)_n$, the product would be polyserine; and if the reading phase were $(-CUU \cdot CUU-)_n$, the expected product would be polyleucine.

A polynucleotide with a repeating tetranucleotide sequence contains four triplets and therefore serves as a template for ~~protein~~ polypeptides with repeating tetrapeptide sequences. Polymers which stimulate incorporation of less than the expected number of amino acids contain terminator ~~triplets~~ codons, such as UAA, ~~and~~ UAG, or ~~the barrier triplet~~ UGA (discussed under Punctuation).

Results obtained with polynucleotides containing repeating sequences of base directly demonstrate nucleotide sequences of terminator codons. Polynucleotides without initiator codons are translated in almost every possible phase and the preferred mode of phasing apparently depends upon the triplet rather

1	2
2	8
3	1
4	5
4+2	3

than the exact distance of the triplet from the 5'-terminus of the polynucleotide. In most cases, base sequences of codons can be derived only if the reading phase of mRNA is known and is correlated with the phase of amino acid incorporation into protein.

The number of codons per amino acid is shown in Table ___. ~~Six degenerate codons correspond to serine, five or six to arginine and to leucine~~ The maximum number of codons for one amino acid is six (serine, leucine and arginine), and from four to one codons for each of the remaining amino acids. It should be noted that only one codon corresponds to tryptophan and one to methionine (codons for N-formyl-methionine, an initiator, are discussed under punctuation).