A NUCLEOTIDE TRIPLET CODE FOR AMINO ACIDS

Carl R. Woese

General Electric Research Laboratory
Schenectady, New York

Received March 27, 1961

We have recently devised a nucleotide triplet code which meets one of the requirements for being the correct solution to the "coding problem" in biology (Woese, 1961). This code was designed to, and does, predict the nucleotide composition of each of the six ribonucleic acid viruses, tobacco mosaic, tomato bushy stunt, southern bean mosaic, turnip yellow, cucumber, and polio, from their particular amino acid compositions. Because the protein and nucleic acid compositions of the six viruses vary considerably, it is probable that the code, while not unique, is one of a relatively small number of closely related codes, all of which fit the experimental data (Yčas, 1960). In the present communication we shall develop this code further.

Nucleotide order within triplets

It is evident that while the composition of a nucleotide triplet corresponding to any given amino acid can be determined by the above mentioned data on virus composition, the nucleotide order within a triplet cannot. This order, however, can be inferred from amino acid replacement data. By this we mean the following. Two similar proteins isolated from different but related species (e.g., Insulin A of ovine vs bovine origin) usually possess identical amino acid sequences except at a few points where one or several amino acids in the one case are "replaced" by different amino acids in the other. These replacements are not random. Using such data, then, we can assume a nucleotide order in any one of the triplets and then arrange the nucleotide order within the remaining triplets so that when one amino acid is "replaced" by another, a minimum number of nucleotide replacements occur in the corresponding triplets. To give an example, ala[1] has been observed to be replaced by leu in some closely related proteins. The nucleotide triplet corresponding to ala is UAG (we assume this to be its order). Leu corresponds to a triplet CGU. If we are to have a minimum number of nucleotide replacements when ala is replaced by leu, we must adjust the order in the leu associated triplet to be UCG. When this is done (see Table 1) for as many amino acids as possible using the aforementioned data, catalogued by Yčas (1958), we arrive at the following ordered nucleotide triplets corresponding to the amino acids: ala-UAG, arg-AGG, asp-GAU, cys-UCC, glu-UAU, gly-GAG, his-UCU, ilu-CAU, leu-UCG, lys-CCG, met-CUU, phe-UUG, pro-CCC, ser-AAG, thr-CAC, try-UUC,

TABLE 1

Amino acid replacements translated into nucleotide triplets

I	II	III	I	II	III
val ↔ ilu CAG CAU	3	1	ala ↔ glu UAG UAG	1	1
ala ↔ gly UAG GAG	3	1	ala ↔ phe UAG UUG	1	1
ala ↔ thr UAG CAC	2	2	ala ↔ tyr UAG UUU	1	2
ala ↔ ser UAG AAG	2	1	glu ↔ asp UAU GAU	1	1
ala ↔ leu UAG UCG	2	1	val ↔ met CAG CUU	1	2
ser ↔ gly AAG GAG	2	1	phe ↔ glu UUG UAU	1	2
val ↔ gly CAG GAG	1	1	arg ↔ lys AGG CCG	1	2
			leu ↔ lys UCG CCG	1	1

I. Amino acid and corresponding nucleotide replacement.
II. Number of occurrences of replacement.
III. Minimum number of nucleotides replaced.

[1] Names of amino acids are abbreviated as in Woese, 1961.

tyr-UUU, and val-CAG. Nucleotide assignments to a few amino acids which occur in small amounts in proteins are as yet somewhat uncertain.

The code applied to other replacement data

As Table 1 shows, most, but not all, of the amino acid replacements correspond to a single nucleotide replacement within a triplet. The exceptions, however, do not invalidate the code, for, among other reasons, there is no proof that a replacement such as ala↔thr did not originally occur as ala←val→thr.

In addition to closely related proteins, which differ in a few amino acids only, it is also possible to recognize distantly related proteins (Gamow *et al*, 1956). In such cases, while most of the amino acid residues are not common to the protein pair, enough of them are, so that some relationship between the two proteins is suggested. Table 2 compares two pairs of distantly related proteins and their corresponding nucleotide triplet sequences. In addition, the region in the adrenocorticotropin molecule which manifests species differences is also included. It is apparent that although single replacements in nucleotide triplets are the majority, many double replacements occur. However, in such a comparison as Insulins A and B, where only 20 per cent of the amino acid residues remain in common, one would expect many amino acids to have undergone at least two replacements. For a random pairing of amino acid sequences one would expect a higher percentage of double and triplet nucleotide replacements, roughly 1:2:3 = 10:16:8. Therefore, we conclude that the present code is consistent with known replacement data.

Chemical mutagens in relation to the present code

Nitrous acid can produce mutations in tobacco mosaic virus ribonucleic acid by oxidizing amino groups on the bases to keto groups (Gierer *et al*, 1958). These changes can then manifest themselves in terms of amino acid replacements. The only fully documented case to date is that of a pro residue being replaced by a leu residue (Tsugita *et al*, 1960). This corresponds on the present code to the nucleotide triplet replacement CCC→UCG.

TABLE 2

Replacement in distantly related proteins

HEMOGLOBIN (Braunitzer *et al*, 1960)

	val	leu	ser	pro	5 ---	ala	asp	lys	thr	10 asp	val
Hbα	CAG	UCG	AAG	CCC	---	UAG	GAU	CCG	CAC	GAU	CAG
Hbβ	CAG	UCU	UCG	CAC	CCC	UAU	UAU	CCG	AAG	UAG	CAG
	val	his	leu	thr	pro	glu	glu	lys	ser	ala	val
	lys	ala	ala	15 try	gly	lys	val	gly	20 ala	his	
Hbα	CCG	UAG	UAG	UUC	GAG	CCG	CAG	GAG	UAG	UCU	
Hbβ	CAC	UAG	UCG	UUC	GAG	CCG	CAG	GAU	CAG	GAU	
	thr	ala	leu	try	gly	lys	val	asp	val	asp	
	ala	gly	glu	25 tyr	gly	ala	glu	ala	30 ser	glu	arg
Hbα	UAG	GAG	UAU	UUU	GAG	UAG	UAU	UAG	AAG	UAU	AGG
Hbβ	UAU	CAG	GAG	---	GAG	UAU	UAG	UCG	GAG	AGG	UCG
	glu	val	gly	---	gly	glu	ala	leu	gly	arg	leu

INSULIN (Sanger, 1960)

	gly	ilu	val	glu	5 glu	lys	cys	ala	ser	10 val	
Ins.A	GAG	CAU	CAG	UAU	UAU	CCG	UCC	UAG	AAG	CAG	
Ins.B	UUG	CAG	GAU	UAU	UCU	UCG	UCC	GAG	AAG	UCU	
	phe	val	asp	glu	his	leu	cys	gly	ser	his	
	cys	ser	leu	tyr	15 glu	leu	glu	asp	tyr	20 cys	asp
Ins.A	UCC	AAG	UCG	UUU	UAU	UCG	UAU	GAU	UUU	UCC	GAU
Ins.B	UCG	CAG	UAU	UAG	UCG	UUU	UCG	CAG	---	UCC	UAU
	leu	val	glu	ala	leu	tyr	leu	val	---	cys	glu

	Hb	Ins.
No. with 0 nucleotide replacements --	9	4
No. with 1 nucleotide replacement ---	13	7
No. with 2 nucleotide replacements --	7	8
No. with 3 nucleotide replacements --	1	1

ADRENOCORTICOTROPIN (Li, 1956; Li *et al*, 1955, 1958, 1961)

	24 pro	gly	ala	glu	asp	asp	30 glu	leu	ala	glu
Pig A	---	GAG	UAG	UAU	GAU	---	---	UCG	---	---
Pig B	---	GAU	GAG	UAG	UAU	---	---	UCG	---	---
	pro	asp	gly	ala	glu	asp	glu	leu	ala	glu
	pro	ala	gly	glu	asp	asp	glu	ala	ser	glu
Sheep	---	UAG	GAG	UAU	GAU	GAU	UAU	UAG	AAG	---
Ox	---	GAU	GAG	UAU	UAG	UAU	GAU	AAG	UAG	---
	pro	asp	gly	glu	ala	glu	asp	ser	ala	glu

One would expect the change C→U, but not C→G on *a priori* grounds. Therefore, we must conclude that unless we postulate some special mechanisms at work in the cell, the present coding scheme is compatible only in part with the chemical mutation data.

The present code vs the commaless code

The present code contains such triplets as CCC and UUU and, consequently, cannot be a commaless code (Crick *et al*, 1957). However, if we exclude the triplets corresponding to the amino aci pro, cys, tyr, arg, met, and try, the remaining twelve triplets belong to a commaless system. In addition, the whole code has the stabilizing constraint that A never appears at the end of a triple Since the present code was not designed to make it commaless, the fact that a portion of it is so, argues in its favor. It might seem, then, that the "true" code is a commaless code, and that the present one is a first approximation to it. We do not hold this view. It will be noticed that the amino acids corresponding to the commaless portion of the present code account for about 80 per cent of the total amino acids in protein. Therefore, the stability of a commaless code would apply to a good portion of the protein synthesis. I call your attention also to the fact that most amino acid replacements occur in proximity to one of the amino acids which corresponds to the "non-commaless' nucleotide triplet (Woese, in preparation).

Stereochemical correlations in the present code

Again, although the code was not designed to show it, there are stereochemical correlations apparent in the present system. For example, gly-GAG, ala-UAG, ser-AAG, val-CAG. Also val-CAG, thr-CAC, and ilu-CAU; or glu-UAU, and asp-GAU; or phe-UUG, and tyr-UUU. These considerations argue in favor of the proposed code Detailed stereochemical discussion will be deferred for the presen

Discussion and Summary

We have shown that there exists a nucleotide triplet code wit the following properties: (1) It predicts the nucleotide composition of six ribonucleic acid viruses from their amino acid composi tion; (2) it is consistent with the known amino acid replacement data; (3) it is partly consistent, at least, with the data on chemically produced mutations; (4) it is in part a commaless code, with an additional stabilizing constraint that one base, A, never appears at one of the positions in a triplet; (5) stereochemically related amino acids correspond to closely related nucleotide triplets. Consequently, we consider this code to be a first approximation to the actual code existing in the living system. The present code appears to have many ramifications in terms of sources of error within cells, mechanisms of synthesis of proteins and nucleic acids, stereochemistry of templates, etc., discussion of which we shall defer until later.

I gratefully acknowledge the encouragement and discussion of Dr. H. M. Rozendaal and the discussions of Doctors J. C. Fisher, C. P. Bean, and W. D. Bellamy. Dr. M. Yčas kindly allowed me to see some of his unpublished tabulations of amino acid sequences, for which I thank him.

Braunitzer, G., Liebold, B., Müller, R., and Rudloff, V. Zeit. Physiol. Chem. 320, 170 (1960).

Crick, F., Griffith, J., and Orgel, L. Proc. Nat. Acad. Sci. U. S. 43, 416 (1957).

Gamow, G., Rich, A., and Yčas, M. Adv. Biol. and Med. Phys. 4, 23 (1956).

Gierer, A., and Mundry, W. Nature 182, 1457 (1958).

Li, C. Adv. in Prot. Chem. 11, 101 (1956).

Li, C., Geschwind, I., Cole, R., Raake, I., Harris, J., and Dixon, J. Nature 176, 687 (1955).

Li, C., Dixon, J., and Chung, D. J. Am. Chem. Soc. 80, 2587 (1958).

Li, C., Dixon, J., and Chung, D. Biochim. et Biophys. Acta 46, 324 (1961).

Sanger, F. Brit. Med. Bull. 16, 183 (1960).

Tsugita, A. and Fraenkel-Conrat, H. Proc. Nat. Acad. Sci. U. S. 46, 636 (1960).

Woese, C. Nature -- submitted for publication (1961).

Yčas, M. in "Symposium on Information Theory in Biology." 70 (1958). Pergamon Press, New York.

Yčas, M. Nature 188, 209 (1960).