S. A. Bernhard*
W. L. Duda

A Note on the Nature of RNA Codes

For some time the "coding problem" for genetics has occupied the attention of biochemists. Briefly, the coding problem is the question of determining the relationships that exist between the sequence of nucleic acids (DNA) in the nucleus of a cell and the amino acids constituting the protein shell outside the nucleus. The predominant assumption concerning this relationship has been that of simplicity. That is, it has been assumed that one could look at a small part of the DNA chain and deduce something about an appropriate area of the protein chain. Recently, Crick et al.[1] have suggested strongly that we might be able to scan DNA chains in chunks of three nucleic acids and deduce the protein chain.

It has further been believed that the coding problem could be restated in terms of a relationship between a sequence of RNA and protein, in which RNA is a sequence of nucleic acids resembling those of DNA. In fact, three of the acids which we name A, C, G are identical in both sets, while DNA contains T, and RNA contains U, as a fourth member. The primary assumption of this relationship is that one of the strands of DNA in the double helix is a map for the construction of an RNA sequence, which in turn is the map for constructing the protein sequence. This DNA-RNA complex has a simple logical relationship operating in a one-to-one fashion. Specifically, with A, C, G, T as the DNA, the relations are as follows: A-U, C-G, G-C, T-A, where the second members of these pairs are elements of the RNA sequence. Thus

A, A, G, T, C, G, T DNA

produces the RNA sequence

U, U, C, A, G, C, A .

Ochoa developed a technique whereby he could synthesize RNA of the type associated with the DNA map. Nirenberg first used this technique to derive the speculation that the code for phenylalanine is UUU. That is, if we looked at the first three letters of the RNA chain and it consisted of UUU, we could then predict that the first letter of the protein sequence was phenylalanine. Similarly, one strand of the DNA chain should be AAA. In quick order, Nirenberg and his co-workers,[2] discovered 15 compositions, but not sequences, of triples which he associated with amino acids. Ochoa and his colleagues[3] determined 19, generally confirming those of Nirenberg. Table 1 gives the RNA code for all 20 amino acids.

During this period, Sueoka[4] undertook a statistical analysis of DNA-protein relationships. In this effort he checked the amino acid distributions for concentrations of C-G DNA which varied from 35% to 72%. The distribution for the 72% concentration is displayed in Table 2.

If we assume the correctness and completeness of the RNA code and work back through the functional relations of protein to DNA, we obtain a prediction of about 50% C-G as the maximum, which sharply contradicts the experimental data of 72%. This leads to the following possibilities:

(1) The RNA code is incorrect.
(2) The RNA code is correct but incomplete.
(3) The hypothesis by Crick, et al. is incorrect.
(4) The theory relating DNA to protein is incorrect.
(5) The DNA-protein code is not universal.

The RNA code may be incorrect. However, in view of the skilled workers who have labored in this area and who have obtained duplicate results, it is important that we have positive reasons for reaching this conclusion.

If we accept possibilities (3) or (4), we can retain the correctness and essential completeness of the RNA code only at the expense of creating a new coding problem between DNA and RNA. Possibility (5) is inappropriate for the present state of biochemistry. While we cannot preclude the possibility of obtaining several large classes of cells such that different coding functions hold within the different classes, it must be remembered that the formulation of the coding problem in the sense of a universal function has served to motivate a remarkable research effort in genetics.[5] For this reason, arguments stressing the possibility of non-universal codes should be treated as speculations until such time as the entertainment of the possibility causes less anguish to theory than the concepts it replaces.

Possibility (2) is of most interest because it retains the RNA code while rejecting the more onerous consequences of the other possibilities. The purpose of

* Institute of Molecular Biology, University of Oregon, Eugene, Oregon.

this Letter is to indicate that the RNA code, together with the Crick hypothesis and the general theory, can be consistent only if there exist RNA coding triples formed from (C, G), e.g., CCC, GGG or some mixture of the two. This is shown as follows:

Assume that every RNA coding triple must contain at least one uracil acid. Suppose now that we wish to maximize the C-G content of the DNA associated with the protein reported by Sueoka. We can do this by associating the 12 triples

ACG	GAC	ACC	AGG
AGC	CAG	CAC	GAG
GCA	CGA	CCA	GGA

uniquely with the 12 most frequent acids. If we assume, further, that Asp-X and Glu-X each contain one acid in negligible quantity we present the most favorable bias toward high C-G content. The C-G content accounted for by these acids is 62%. This leaves us with five acids that can contribute at most $\frac{1}{3}$ of their mixture to the C-G content. These acids are Cys, Met, Lys, His and Phe, and the maximum contribution of C-G is 3%. Consequently, the maximum C-G content for the DNA associated with this protein is 65%. If we assume that tryptophan is correctly coded by UGG, the maximum C-G content drops below 64%. Similarly if Glu-X and Asp-X each contain two acids in significant proportions, the maximum C-G content would drop more sharply. This deviation of predicted C-G content from experimentally obtained C-G contradicts the assumption of a uracil acid in each RNA coding triple.[6]

We ignore now the question of amino acid distribution in protein and consider the C-G content of 74% mentioned by Sueoka.[4] If we assume that every RNA coding triple must contain at least one U or A, then the maximum C-G content of the corresponding DNA must be 67%, which deviates from the experimentally determined content by 7%. Consequently, we deduce that there must exist RNA codes that contain neither A nor U or, positively, that there exist RNA code triples composed solely from C and G.

Summary

From an analysis of the RNA code and known distributions of protein and C-G content, it is shown that:

(1) There exist RNA coding triples containing no U or A; or
(2) The RNA code is incorrect; or
(3) The code relating DNA to RNA is different from that generally supposed.

Table 1 **Amino acid code.***

Amino Acids	*RNA Bases*†
Phenylalanine	UUU
Alanine	UCG
Arginine	UCG
Aspartic Acid	UAG
Asparagine	UAA, UAC
Cysteine	UUG
Glutamic Acid	UAG
Glutamine	UCG‡
Glycine	UGG
Histidine	UAC
Isoleucine	UUA
Leucine	UUC, UUG, UUA
Lysine	UAA
Methionine	UAG
Proline	UCC
Serine	UUC
Threonine	UAC, UCC
Tryptophan	UGG
Tyrosine	UUA
Valine	UUG

* As cited in *The New York Times*, February 2, 1962. [*Note added in proof*: See also p. 443 in the paper by Speyer et al[3].]
† U = uracil; C = cytosine; G = guanine; A = adenine.
‡ Predicted, no experimental evidence.
Base triplets not containing U may exist.
Each sequence of bases has not yet been determined, which explains why some appear with more than one amino acid.

Table 2 **Amino acid distribution for 72% concentration of C-G DNA.****

	72% C-G *Stable*
Lys	5.5
His	2.1
Arg	6.5
Asp-X	12.4
Glu-X	17.5
Pro	6.2
Ala	21.6
Val	11.8
Leu	11.2
Tyr	2.1
Phe	3.0
	Unstable
Gly	15.2
Thr	8.1
Ser	5.8
Ileu	5.0
Met	2.1
Cys	—

** The stable and unstable acids are expressed as a percentage of the total amount of stable acid. [The correction factor for absolute percentage is 0.734.]

References and footnotes

1. F. H. Crick, L. Barnett, S. Brenner and R. J. Watts-Tobin, *Nature* **192**, 1227 (December 30, 1961).
2. J. H. Matthaei, O. W. Jones, R. G. Martin and M. W. Nirenberg, *Proc. Nat. Acad. Sci.* **48**, 666 (1962).
3. J. F. Speyer and Peter Lenglye, Carlos Bosilio and S. Ochoa, *Proc. Nat. Acad. Sci.* **48**, 441 (1962).
4. N. Sueoka, *Proc. Nat. Acad. Sci.* **47**, 1141 (1961).

5. We might note that although the notion of a universal function (or code) can never be completely proven, it can be disproven.
6. The argument against all triples containing uracil is also contained in the argument which is based solely on C-G content. However, to strengthen the case against possible experimental errors, the argument from protein was used to give a wider deviation of prediction from experimental fact.

Received March 23, 1962