Det Kongelige Danske Videnskabernes Selskab
Biologiske Meddelelser, bind **22**, nr. 8

Dan. Biol. Medd. **22**, no. 8 (1955)

DEDICATED TO PROFESSOR NIELS BOHR ON THE OCCASION OF HIS 70TH BIRTHDAY

ON INFORMATION TRANSFER FROM NUCLEIC ACIDS TO PROTEINS

BY

G. GAMOW

København 1955
i kommission hos Ejnar Munksgaard

Printed in Denmark
Bianco Lunos Bogtrykkeri A-S

One of the most important problems of today's theoretical biology is the problem of the transfer of the hereditary informations from the chromosomes, where they are presumably stored in the form of long polynucleotide sequences, to the enzymes which, as all other proteins, are represented by long polypeptide chains. Mathematically the problem reduces to finding a procedure by which a long number written in fourdigital system (four bases forming the molecules of nucleic acid) can be translated in a unique way into a long word formed by twenty different letters (twenty amino acids which form protein molecules). The fact that number 20 represents the number of different triplets (with disregard to order) which can be formed out of 4 different elements suggests that each amino acid in the resulting polypeptide sequence is determined by a group of three bases in the corresponding section of polynucleotide chain. Some time ago the author attempted to establish such a translation mechanism on the basis of Crick and Watson's model of Deoxyribonucleic Acid (DNA) molecule[1]. It turned out, however, that the translation mechanism suggested by DNA structure ("diamond code") leads to a contradiction with the actually observed sequences of amino acids in certain protein molecules. This negative result is presumably due to over-simplification of the original picture, since it seems, indeed, that the transfer of informations from chromosomes to enzymes is a two-step process. First the informations stored in DNA molecules are transmitted to the molecules of RNA (Ribonucleic Acid) which move out into the cytoplasm of the cell and form the so-called microsomes. The second part

[1] G. Gamow. Dan. Biol. Medd. 22, no. 3 (1954).

of the process consists in synthesizing the proteins according to the informations carried by RNA-base sequences. The presence of such double coding makes the analysis of information transfer much more difficult.

The most promising biological material for the study of information transfer is presented by viruses, since it has been shown that in this case (at least for bacterial viruses, or phages) the virus protein for the progeny is synthesized directly by the nucleic acid of the original virus particles which had penetrated into the cytoplasm of the host cell[1]. If we limit our studies to plant viruses in which nucleic acid is of RNA-type (rather than DNA-type, as in all animal viruses or phages), we may also hope that we are dealing with only one-step information transfer.

Very careful analysis of both nucleic acid and protein constitution is available in the case of Tobacco Mosaic Virus[2], and is shown in Tables I and II.

TABLE I.

Relative amounts of bases (in moles) in RNA from TMV.

	1. Adenine	2. Guanine	3. Cytosine	4. Urosil
TMV	0.31	0.25	0.18	0.26

TABLE II.

Relative amounts of amino acids (in moles) in proteins from TMV.

Amino Acid	TMV	Amino Acid	TMV
1. Alanine	0.070	11. Leucine	0.087
2. Arginine	0.069	12. Lysine	0.012
3. Asp. acid	0.030	13. Methionine	0.000
4. Asparagine	0.090	14. Phenylalanine	0.062
5. Cysteine	0.010	15. Proline	0.061
6. Glu. acid	0.047	16. Serine	0.083
7. Glutamine	0.047	17. Threonine	0.102
8. Glycine	0.031	18. Tryptophane	0.012
9. Histidine	0.000	19. Tyrosine	0.026
10. Isoleucine	0.061	20. Valine	0.097

[1] A. D. HARSHEY and M. CHASE. J. Gen. Physiol. **36**, 39 (1952).
[2] C. A. KNIGHT. J. Biol. Chem. **171**, 297 (1947); **197**, 241 (1952).

If we assume the hypothesis that each amino acid in the protein structure is determined by a triplet of bases in RNA chain, we may expect that relative abundance of different amino acids

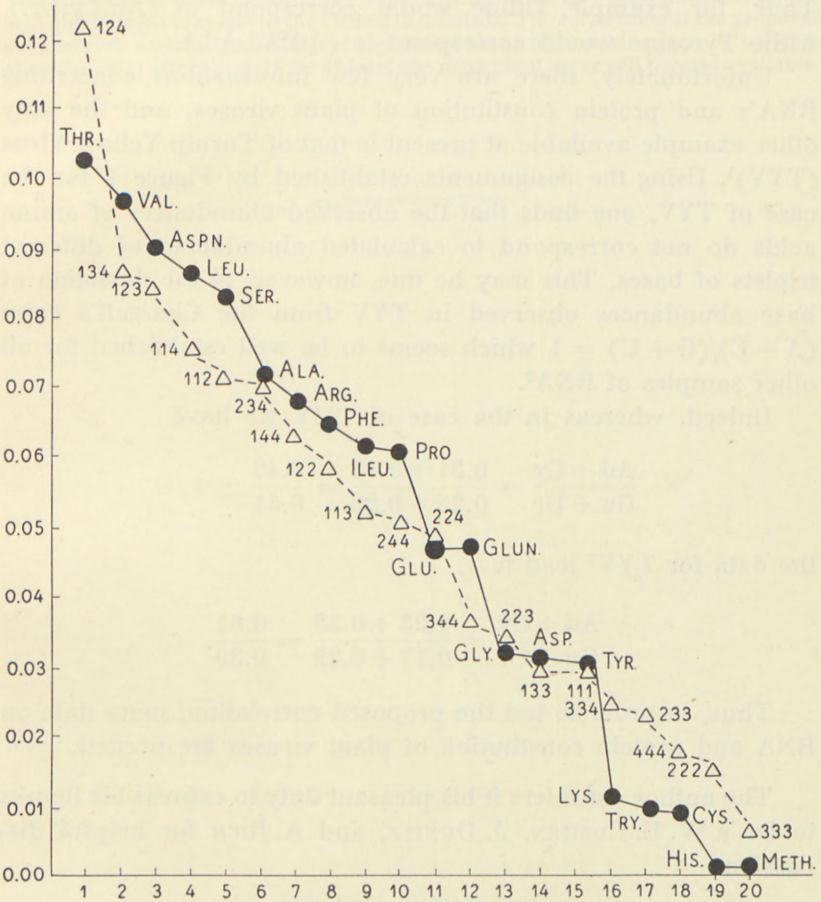

Fig. 1

will be given by the products of relative abundances of the three corresponding bases. (Because of multiplicity of various types of triplets the probability of *aaa*, *aab*, and *abc* types of triplets should be given the weights 1, 3, and 6).

The calculated abundance of various triplets, and the observed abundances of various amino acids, are plotted in Figure 1

in decreasing order, and we see that two sets of data are in a reasonably good agreement. If this agreement is not coincidental, it should permit the establishment of a correlation between individual units forming polynucleotide and polypeptide chains. Thus, for example, Valine would correspond to (Ad.Cy.Ur.), while Tyrosine would correspond to (Ad.Ad.Ad.).

Unfortunately, there are very few informations concerning RNA's and protein constitution of plant viruses, and the only other example available at present is that of Turnip Yellow Virus (TYV)[1]. Using the assignments established by Figure 1 for the case of TYV, one finds that the observed abundances of amino acids do not correspond to calculated abundances of different triplets of bases. This may be due, however, to the deviation of base abundances observed in TYV from the Chargaff's rule: $(A + C)/(G + U) = 1$ which seems to be well established for all other samples of RNA[2].

Indeed, whereas in the case of TMV we have

$$\frac{Ad + Cy}{Gu + Ur} = \frac{0.31 + 0.18}{0.25 + 0.26} = \frac{0.49}{0.51} \simeq 1,$$

the data for TYV[2] lead to

$$\frac{Ad + Cy}{Gu + Ur} = \frac{0.23 + 0.38}{0.17 + 0.22} = \frac{0.61}{0.39}.$$

Thus, in order to test the proposed correlation, more data on RNA and protein constitution of plant viruses are needed.

The author considers it his pleasant duty to express his thanks to Dr.'s W. E. Cushen, J. Dunitz, and A. Rich for helpful discussion.

Note added in April 1955.

It was indicated to the author by his colleague Dr. M. Yčas that a reasonable agreement between amino acid constitution of *both* viruses (TMV and TYV), and the probabilities of base-triplets can be achieved if the "burden of fit" is distributed equally between the two caces. Indeed, if one makes the following assignments of the most abundant amino acids to the most probable base-triplets:

[1] R. Markham and J. D. Smith. Bioch. J. 49, 401. (1951); E. Roberts and G. B. Ramasarma. Proc. Soc. Exp. Biol. Med. 80, 101 (1952).

[2] D. Elson and E. Chargaff. Biochim. et Biophys. Acta (in press). The author is grateful to Dr. Chargaff for the opportunity to see this manuscript prior to publication.

Ala → 113; Arg → 122; Asp + Aspn → 124 + 222 (?); Glu + Glun → 144 + 223; Gly → 344; Jleu → 133; Leu → 123; Lys → 233; Phe → 244; Pro → 334; Ser → 234; Thr → 134; Vol → 114,

the observed abundances of these amino acids in both TMV and TYV agree with the calculated probabilities of base-triplets with a mean error of 20 per cent. This error can be easily explained by experimental errors in measurements of amino acid and base percentages in the viruses in question. The correctness of the proposed assignement, and the possibility of any assignement of that kind, can be verified, of course, only after the data for at least one more plant virus will become available.

Indleveret til selskabet den 28. marts 1955.
Færdig fra trykkeriet den 19. september 1955.