Det Kongelige Danske Videnskabernes Selskab
Biologiske Meddelelser, bind 22, nr. 3

Dan. Biol. Medd. 22, no. 3 (1954)

POSSIBLE MATHEMATICAL RELATION BETWEEN DEOXYRIBONUCLEIC ACID AND PROTEINS

BY

G. GAMOW

København
i kommission hos Ejnar Munksgaard
1954

DET KONGELIGE DANSKE VIDENSKABERNES SELSKAB udgiver følgende publikationsrækker:

L'Académie Royale des Sciences et des Lettres de Danemark publie les séries suivantes:

	Bibliografisk forkortelse *Abréviation bibliographique*
Oversigt over selskabets virksomhed (8°) *(Annuaire)*	Dan. Vid. Selsk. Overs.
Historisk-filologiske Meddelelser (8°)	Dan. Hist. Filol. Medd.
Historisk-filologiske Skrifter (4°) *(Histoire et Philologie)*	Dan. Hist. Filol. Skr.
Arkæologisk-kunsthistoriske Meddelelser (8°)	Dan. Arkæol. Kunsthist. Medd.
Arkæologisk-kunsthistoriske Skrifter (4°) *(Archéologie et Histoire de l'Art)*	Dan. Arkæol. Kunsthist. Skr.
Filosofiske Meddelelser (8°) *(Philosophie)*	Dan. Filos. Medd.
Matematisk-fysiske Meddelelser (8°) *(Mathématiques et Physique)*	Dan. Mat. Fys. Medd.
Biologiske Meddelelser (8°)	Dan. Biol. Medd.
Biologiske Skrifter (4°) *(Biologie)*	Dan. Biol. Skr.

Selskabets sekretariat og postadresse: Ny vestergade 23, København V.

L'adresse postale du secrétariat de l'Académie est:

Det Kongelige Danske Videnskabernes Selskab,
Ny vestergade 23, Copenhague V, Danemark.

Selskabets kommissionær: EJNAR MUNKSGAARD's forlag, Nørregade 6, København K.

Les publications sont en vente chez le commissionnaire:
EJNAR MUNKSGAARD, *éditeur, Nørregade 6, Copenhague K, Danemark.*

Det Kongelige Danske Videnskabernes Selskab
Biologiske Meddelelser, bind **22**, nr. 3

Dan. Biol. Medd. **22**, no. 3 (1954)

POSSIBLE MATHEMATICAL RELATION BETWEEN DEOXYRIBONUCLEIC ACID AND PROTEINS

BY

G. GAMOW

København
i kommission hos Ejnar Munksgaard
1954

Printed in Denmark
Bianco Lunos Bogtrykkeri A-S

»Die Methoden der Polypeptidsynthese gestatten den Aufbau langer Ketten mit vielfachen Variationen in der Reihenfolge. Es ist drum kein blosses Spiel mit Zahlen, wenn man die gegebenen Möglichkeiten berechnet.«

(EMIL FISCHER, Sitzungsber. der Kgl. Preuss. Akad. der Wiss., p. 990, 1916.)

It was recently shown by WATSON and CRICK (1) that the molecules of *Deoxyribonucleic Acid* (DNA), which constitute chromosome fibers of living cells, are formed by double sequences of four basic compounds (*Adenine, Thymine, Guanine,* and *Cytosine*) held together by two parallel sugar-phosphate chains. Since, according to that scheme, *Adenine* can pair only with *Thymine,* and *Guanine* only with *Cytosine,* one half of that double sequence of bases is completely determined by the other half. Thus, if we denote the four bases by figures 1, 2, 3, 4 (a mathematician would prefer 0, 1, 2, 3), all hereditary properties of any living organism should be characterized by *a long number* ("number of the beast") *written in a four digital system,* and containing many thousands of consecutive digits. The numbers describing two different members of the same species must be very similar to each other (though not quite identical, unless they belong to a pair of identical twins), whereas the numbers representing the members of two different species must show larger differences. Since the number of all possible arrangements of four elements in sequences of several thousand is incredibly large*, we must conclude that all living organisms represent only a negligible fraction of all "mathematically possible" forms of life. For example, it is extremely unlikely that any organism which ever lived on the surface of the earth was represented by such familiar numbers as π, or $\sqrt{3}$ written in the four digital system. WATSON and CRICK

* For a chain which is, for example, 10,000 steps long, the number of possible arrangements is $4^{10,000} = 10^{6,000}$, which is much (much!) larger than the number 10^{76} representing the number of all atoms in the Universe within range of the 200″ telescope at Palomar Mountain Observatory.

TABLE I. The list of Amino acids.

1. *Glu*tamic acid	14. *Phe*nylalanine
2. *Leu*cine	15. *Cys*tine
3. *Asp*artic acid	16. *His*tidine
4. *Ser*ine	17. *Meth*ionine
5. *Lys*ine	18. *Tryp*tophane
6. *Gly*cine	19. *Cyst*eic acid
7. *Val*ine	20. *Hydr*oxyproline
8. *Pro*line	21. (Norvaline)
9. *Arg*inine	22. (Hydroxy glutamic acid)
10. *Ala*nine	23. (Asparagine)
11. *Thr*eonine	24. (Glutamine)
12. *Iso*leucine	25. (Cannine)
13. *Tyro*sine	

also suggest a very plausible mechanism by which the replication of DNA molecules may take place, with a provision for occasional mutations, i. e. the change of some digits in the original "number of the beast".

It is well known, however, that, while the chromosomes are responsible for carrying all (or, at least, most of) the hereditary information from the parents to the progeny, the actual work of the growing and the development of any organism are carried out by enzymes which catalyze various biochemical reactions in the cytoplasm of the cell. In fact, while a chromosome can be compared with the file cabinet in the director's office of a large factory where all the blueprints are stored, the enzymes play the role of engineers, foreman, and workers constituting the major part of the entire outfit. It follows that there must exist a very precise mechanism which shapes the enzymes produced in any given organism, exactly according to the hereditary information carried by chromosomes. The enzymes, being proteins, possess, however, an entirely different constitution than DNA molecules, and are formed by long sequences of many different amino acids. The number of different amino acids which participates in the structure of proteins is usually taken as 20, although actually there may be a few more. In Table I these amino acids are listed in order of their relative abundance in proteins.

If one assigns a letter of the alphabet to each amino acid, each protein (and, in particular, each enzyme) can be considered

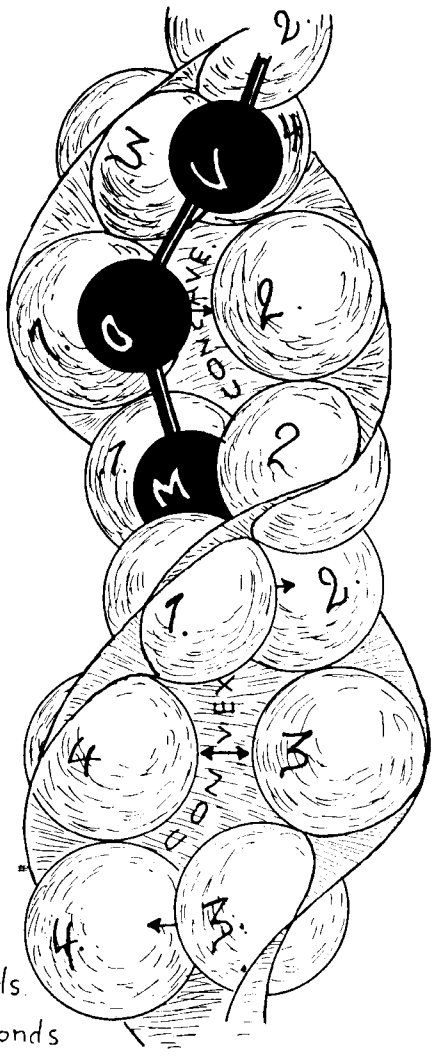


Fig 1.

White balls: bases.
Black balls: amino acids.
Arrows: H bonds
Bars: Peptide bonds

as *a long word based on an alphabet with 20* (or somewhat more) *different letters.*

It was recently suggested by the author (2) that a simple relation between the order of bases in chromosome fibers, and the

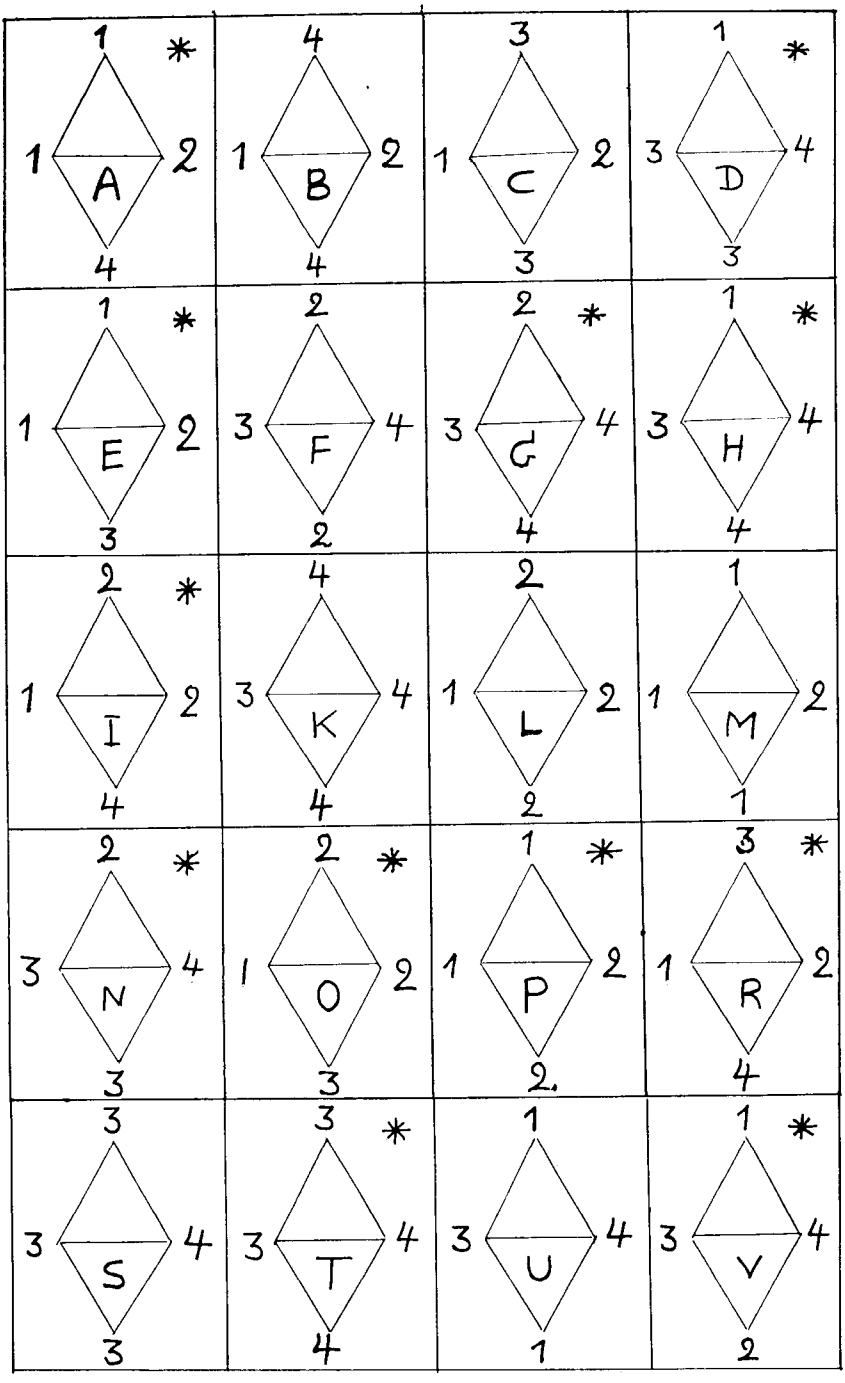
order of amino acids in the corresponding enzymes, can be established by considering basic geometrical features of the WATSON and CRICK model of a DNA molecule.

In fact, as can be easily seen from Fig. 1, the helical nature of the DNA molecule gives rise to diamond-shaped configurations formed by four bases. The two bases at both ends of the horizontal diagonal can be only Adenine-Thymine or Guanine-Cytosine pairs, whereas the upper and lower corners of each "diamond" may be occupied by any of the four bases. Thus, the total number of different "diamonds" is given by the number of different triple combinations of four elements, and one can easily calculate that it is equal to *20*. These 20 different "diamonds" are shown in Table II, where the numerals 1, 2, 3, and 4 stand for Adenine, Thymine, Guanine, and Cytosine, respectively. Each type of "diamond" is denoted by a letter of the alphabet*. The idea is that the amino acids participating in the structure of proteins are those whose residues fit into the various diamond-shaped cavities formed by the combinations of four bases in the DNA molecule. Since the sequence of bases determines in a unique way the sequence of diamonds, we have here a mechanism for the production of a specific set of enzymes by each particular chromosome. According to this point of view, DNA molecules act as highly specialized catalysts, arranging the amino acids from the surrounding medium in well defined sequences, and holding them in that position long enough to permit the amino group of each of them to combine with the hydroxyl group of its next neighbour.

The proposed scheme is quite consistent with the existing data on the linear dimensions of polynucleotide and polypeptide chains. In fact, the distance between the P-atoms in the DNA molecule (along the spiral line) is about 7Å, while the distance between two C_α-atoms in an extended polypeptide chain is 3.6 Å. If one tries to construct a polypeptide chain on the *convex* side of the DNA helix, one gets two amino acids for each "diamond" in the polynucleotide chain, which is certainly not correct. However, constructing the same chain on the *concave* side of the DNA helix, one gets the correct relationship of one amino acid for each dia-

* The assignment of letters in Fig. 2 is different from that given in the author's original article (Ref. 2). The new assignment associates more probable diamonds with more abundant letters of the alphabet in the English text, which makes the "words" made up by various diamond sequences more pronounceable.

TABLE II.


1 * 1 A 2 4
4 1 B 2 4
3 1 C 2 3
1 * 3 D 4 3
1 * 1 E 2 3
2 3 F 4 2
2 * 3 G 4 4
1 * 3 H 4 4
2 * 1 I 2 4
4 3 K 4 4
2 1 L 2 2
1 1 M 2 1
2 * 3 N 4 3
2 * 1 O 2 3
1 * 1 P 2 2.
3 * 1 R 2 4
3 3 S 4 3
3 * 3 T 4 4
1 3 U 4 1
1 * 3 V 4 2

mond-shaped cavity (3.) Thus, we may conclude that *enzyme molecules grow "on the inside" of the chromosome helix*, as indicated schematically in Fig. 1.

Inspecting the 20 diamonds shown in Table II, one notices that 12 of them (marked by asterisks) can exist in two bilaterally symmetrical forms. Thus, for example, the letter *H*, which is given in the table as $\left(3 \cdot \frac{1}{4} \cdot 4\right)$, can also exist in the form $\left(4 \cdot \frac{1}{4} \cdot 3\right)$. If one considers the two bilaterally symmetrical forms as two different entities, one should raise the number of different diamonds to 32. However, inspection of the structural formulae of amino acids indicates that, with possibly a few exceptions, the bilateral symmetry of the diamonds is of no importance. In fact, there are only three or four amino acids whose residues possess no bilateral

TABLE III.
Possible combination of "diamonds".

A, E, I, O associates with A, D, E, F, G, H, I, L, M, N, O, P, U, V
D, G, H, N associates with A, B, C, D, E, G, H, I, K, N, O, R, S, T
L, M, P associates with A, E, I, O, L, M, P
K, S, T associates with D, G, H, K, N, S, T
B, C, R associates with D, F, G, H, N, U, V
F, U, V associates with A, B, C, E, I, O, R

symmetry around the C_α-bond. The existence of these particular amino acids may account for the "excess over 20" of the total number of amino acids which can be used by DNA molecules for the process of protein-(enzyme)-synthesis.

If the theory of DNA protein correlation described on the previous pages is correct, there must exist a way of establishing a unique correspondence between the amino acids, given in Table I, and the diamonds given in Table II. This correspondence should be based on the expected intersymbol correlation between various diamond-shaped cavities provided by the polynucleotide helix, and the observed intersymbol correlation between the amino acids as arranged in polypeptide chains.

The former correlation can be easily established by considering various possible combinations of the diamonds listed in the table. Thus we find, for example, that the letter A can be associated with the letters M, N, G, etc., but cannot be associated with the letters B, C, R, etc. The complete list of possible "pairs of diamonds" is given in Table III.

We notice that eight diamonds (A, D, E, G, H, I, N, O) have a much larger "affinity" for other diamonds than the remaining twelve. The table also shows that the six letters (B, C, F, R, U, V) can occur only singly, the twelve letters (A, D, E, G, H, I, K, L, M, N, O, S) can occur in pairs, and the remaining two (P and T) can repeat consecutively an unlimited number of times.

The empirical information on the order in which various amino acids occur in protein molecules is, however, very meager. The only protein, for which the sequence is known, is insulin which was studied by F. SANGER and his collaborators (4). The "words" representing two insulin chains, known as A and B, are:

Gly-Isol-Val-Glu-Glu-Cy-Cy-Ala-Ser-Val-Cy-Ser-Leu-Tyr-Glu-Leu-Ast-Tyr-Cy-Asp

and

Phe-Val-Asp-Glu-His-Leu-Cy-Gly-Ser-His-Leu-Val-Glu-Ala-Leu-Tyr-Leu-Val-Cy-Gly-Glu-Arg-Gly-Phe-Phe-Tyr-Thr-Pro-Lys-Ala.

The two series are *21* and *30* terms long, respectively; they make use of 16 different amino acids listed in Table I.

Any attempt to establish a one-to-one correspondence between twenty amino acids of Table I and twenty "diamonds" of

Table II must face the fact that the number of possible assignments is immensely large, being given by: $20! = 2.3 \cdot 10^{17}$. However, due to highly restricting intersymbol correlation rules given in Table III, such an attempt becomes possible. Thus, for example, there are only sixteen different possible cases in which a double letter is followed by another double letter. These sixteen cases break up into four groups in such a way that the members

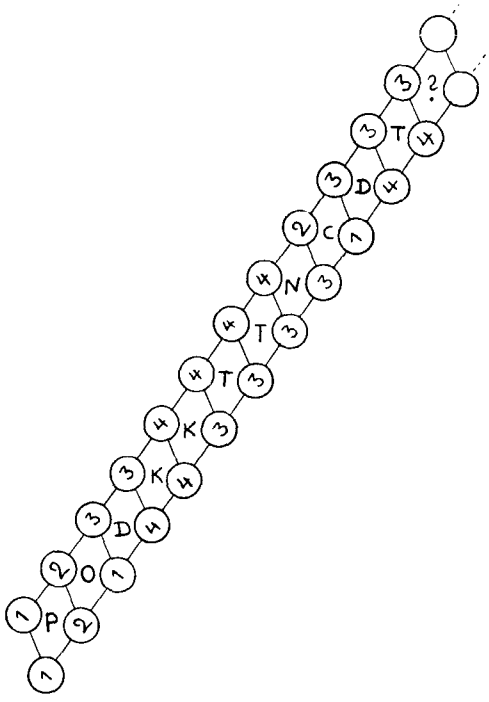


Fig. 2.

of the same group can be transformed into each other by a simple internal permutation. Thus, in order to explain the sequence:

Glu-Glu-Cy-Cy

in insulin A, one must try only four different possibilities (A-A-N-N; L-L-M-M; K-K-T-T, and T-T-S-S). In each of these four cases, we also know the letter-assignment of the fourth amino acid to the right from Cy-Cy, since it is also a Cy. This restricts the choice of the letter for Val since it must precede both Glu and Cy.

Proceeding in this way, it is possible to "decipher" the first eleven places in insulin A as:

P-O-D-K-K-T-T-N-C-D-T- (cf. Fig. 2).

The choice of nine letters from the first K to the last T is unique, except for the above-mentioned internal permutations. However, the next step runs into a difficulty. The 12th amino acid is *Ser*, which was already assigned the letter *C*. But, according to Table III, the letter *C* cannot follow the letter *T*. Thus, at least in this particular case, the correspondence between amino acids and "diamonds" cannot be established, although one comes rather close to it.

It is possible that the difficulty encountered in "deciphering" the structure of insulin in terms of "diamonds" is due to oversimplification of the situation in the proposed form of the theory. For example, it is not impossible that some of the "diamonds" can accommodate more than one amino acid, or that some amino acid can fit into more than one diamond. The answer can be given only by actually constructing a model of a DNA molecule from an atomic-model-kit, and comparing it with the models of various amino acids.

But it looks more likely that insulin is not a good case for testing the validity of the proposed theory. In fact, the theory pertains primarily to "hereditary proteins", i. e. the proteins whose structure is directly and completely determined by genes in the chromosomes. Thus, for example, it should apply directly to the substances primarily responsible for colour vision, or coagulation of blood which are subject to strict laws of heredity. It is not at all certain that insulin falls into this category of organic proteins, especially because diabetes, a sickness connected with insulin deficiency, does not seem to possess hereditary characteristics.

Unfortunately, however, we have no information concerning amino acid sequence in any other protein.

A somewhat different approach to the problem can be provided by the study of relative abundances of different amino acids in various proteins. One would, in fact, expect that different types of "diamonds" are affected in different ways by the variation of *Adenine-Thymine* to *Guanine-Cytosino* ratio. Suppose that, within a particular group of living organisms, the abundance of the first pair of bases varies between the limits of $(X - \Delta X)$ and

$(X + \Delta X)$, so that the abundance of the second pair lies within the limits $(1 - X \pm \Delta X)$. It is easy to see that, in such a case, the expected variability of different types of "diamonds" will be different. The "diamonds" defined entirely by the numbers of the first pair (α-type), such as, for example, $\left(1\frac{1}{1}2\right)$ or $\left(1\frac{1}{2}2\right)$, will appear with the relative probability X^3 (since there are only three free choices), whereas the diamonds of the type $\left(3\frac{3}{3}4\right)$ or $\left(3\frac{3}{4}4\right)$ (δ-type) will have the relative probability $(1 - X)^3$. The dia-

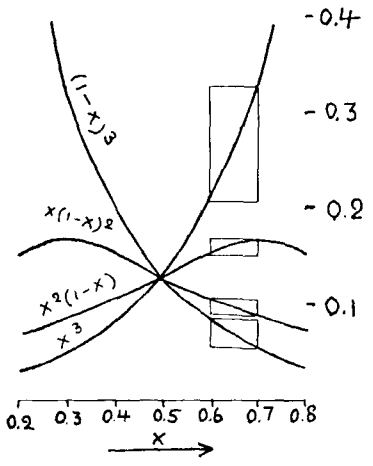


Fig. 3.

monds of the *mixed* type, such as $\left(1\frac{1}{4}2\right)$ (β-type), with the excess of first pair of bases, or such as $\left(3\frac{2}{3}4\right)$ (γ-type) with the excess of second pair of bases, will have relative probabilities given by $X^2(1 - X)$, and $X(1 - X)^2$, respectively. These four functions are plotted in Fig. 3 in respect of X. We notice that, in case that X varies, let us say, within the limits 0.65 ± 0.05, the relative probability of α-type diamonds varies in rather wide limits, whereas the probability of β-type diamonds remains almost constant because the corresponding curve passes through a maximum near $X = 0.65$. This theoretical result may be correlated with observations (5) which seem to suggest that some of the amino

acids show wider variations in different proteins than some others. It would be interesting to undertake a special investigation by comparing the relative abundances of various amino acids in the cytoplasm of cells with the ratios of base-pairs in the corresponding nuclei.

It is the author's pleasant duty to express thanks to Dr. F. H. C. CRICK, and also to the members of the Cyclotron Laboratory (D.T. M.) of the Carnegie Institution of Washington, for helpful discussion of the problems.

Note added in the proof (June, 1954):

Since this article was sent to print, several alternative attempts were made to decipher the sequences of amino-acids in protein molecules in terms of base-sequences in the molecules of nucleic acids. Although no finite solution of that basic problem has as yet been found, a number of interesting possible relationships came to light. The work in this direction is now being continued, and will be reported in due time.

The George Washington University.
Washington, D. C.
U. S. A.

References.

(1) I. D. WATSON and F. H. C. CRICK, Nature **171,** 964 (1953).
(2) G. GAMOW, Nature **173,** 318 (1954).
(3) This remark is due DR. F. H. C. CRICK.
(4) F. SANGER and H. TUPPY, Biochem. Journal, **49,** 481 (1951); F. SANGER and O. P. THOMPSON, Biochem. Journal, **53,** 336 (1953).
(5) Information theory in Biology. Edited by H. QUASTLER, University of Illinois Press (1953).

Indleveret til selskabet den 5. februar 1954.
Færdig fra trykkeriet den 4. september 1954.