Biological Dictionary

Okan Gurel

(Z. Naturforsch. **30 c**, 843–846 [1975]; received July 14, 1975)

α-Helix, Amino Acids, Proteins, Hexagonal Neighborhood, Biological Communication

It is proposed that the biological communication is a language process with the basic word structure formed as a hexagonal neighborhood on the α-helix region of protein molecules.

The commonly accepted view is that the tertiary structure of proteins plays an important role in their biological activities. Studying the tertiary structure should, therefore, help us to understand this role. Many investigations have been directed to relating the primary structure (the amino acid sequence) of proteins to the tertiary structure. Algorithms relating primary structure to tertiary structure have been suggested, however it is clear that the analysis of isolated amino acid positions do not yield any clue to our understanding of the biological activity. A group of amino acids has also been considered in some other studies. By considering the triplet of amino acids; a considerable success has been recorded by Wu and Kabat [1] reported in a series of papers appearing since 1971. The tertiary structure prediction via triplets appears to be promising. An extensive study based on pairs of amino acids has also been reported by S. Erhan and his associates [2].

The state of art is summarized in Hruby [3]: "Little has yet been learned in any predictive way about the relationship of conformation to the biological activity of small peptides, but this is, perhaps, as much due to our lack in understanding the chemistry of the biological activity of peptides, as well as to our ignorance about the significant conformational factors related to the biological activity."

Just as in the case of the bridge between DNA and amino acids, there is a close relationship between the amino acids and the function they perform in the form of proteins. The first bridge is completely described by the genetic code, however the second bridge is still missing. The basic difficulty stems from the fact that the first one is a *decoding* process, the second is a complete communication process, a *language*.

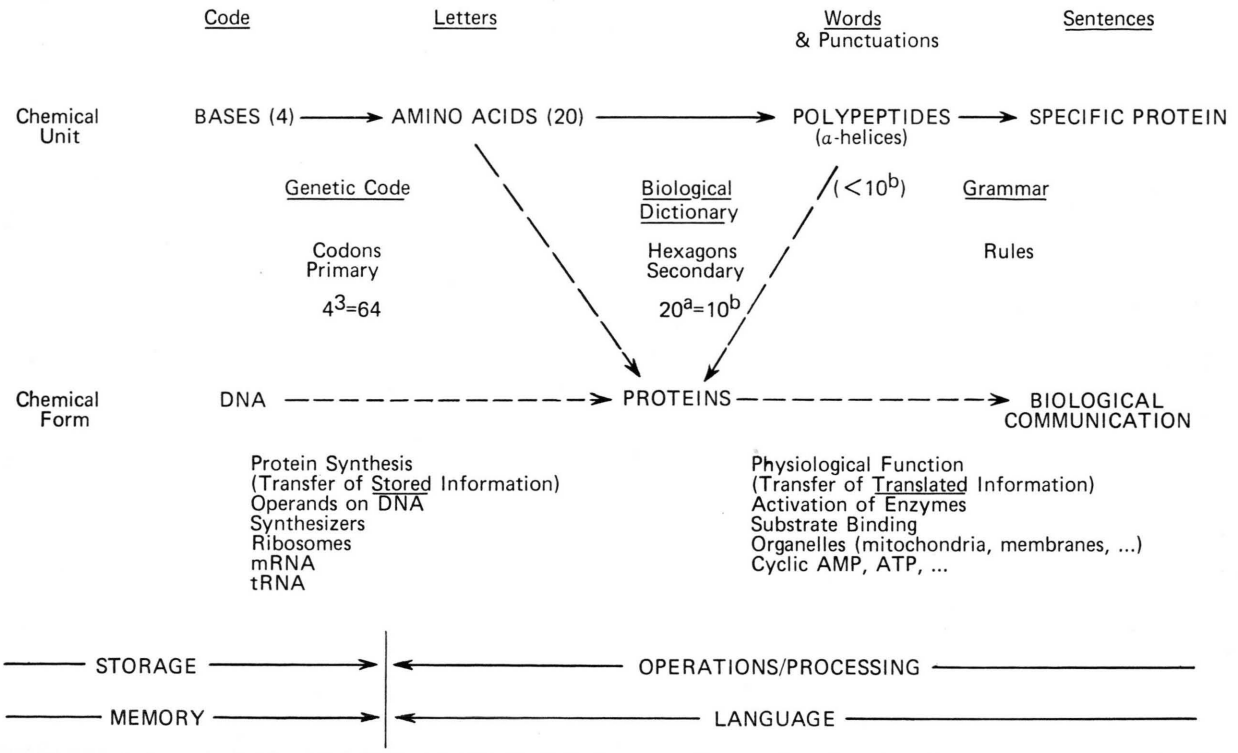

Fig. 1. A diagram showing the relation of biological dictionary to the genetic code and to the grammar necessary for biological communication at the protein level.

Requests for reprints should be sent to O. Gurel, IBM International Business Machines Corp., 1133 Westchester Avenue, *White Plains*, N. Y., 10604, USA.

As illustrated schematically in Fig. 1, the set of amino acids form the *alphabet* of life. The DNA bases conveniently reduce the necessary information for the amino acids alphabet to a compact form. Between the alphabet and the actual message carried by proteins to perform certain physiological/metabolic activity which can be viewed as a *sentence*, there must exist a logical set of *words*, namely a dictionary, from which sentences of the biological *language* can be formed.

With the above basic view in mind, we can see that knowing the alphabet, understanding the function (tertiary structure) may not be possible unless some intelligent subdivision is introduced in the form of "words". This should also reveal some grammatical rules, however, requiring examination of sentence structures (proteins and their functions).

In the present communication *the word structure for this language is proposed.* There exist some indirect attempts in this direction, however the basic philosophy of these investigations differ from that in the present proposal. For example, Wu and Kabat mimic the codon of genetic code in taking tripeptides [1a and c] in relation to helical regions (secondary structure) and study "the influence of nearest-neighbor amino acids on the confirmation of the middle amino acid in proteins". In Kabat and Wu [1b] a significant observation is reminded to the reader by referring to earlier works of others: "Different sets of values indicate the influence of other factors besides nearest neighbors in three dimensional structure, as is well known. Thus, helix formation is influenced by residues at positions $(n-4)$, $(n-3)$, $(n+3)$. $(n+4)$, and perturbing effects of other factors such as contact with, or proximity to, a prosthetic group may also be substantial".

It is proposed here that the secret lies in L. Pauling's [4] α-helix which was indirectly a corner stone also in the prediction of the double helix of DNA in 1953. The characteristic feature of the α-helix of the protein backbone is revealed, if the cylindrical surface of the helix is slit open in parallel to the axis of the cylinder, Fig. 2. If an amino acid is located say (1), the neighboring amino acids are not only 2 and 3 (if $n =$ "1", "2" $= n-1$, "3" $= n+1$), but also 4, 5, 6, and (7). This *hexagon* is the basic unit of the surface created by the α-helix. Before elaborating further on this hexagon, we can observe that there are, however artificial, two more layers on the α-helix: i) 8, 9, ..., 12, 13, and ii) 14, 15, ..., 18, 19. These are not only consecutive layers on the α-helix but also equivalent to layers of the three concentric hexagons

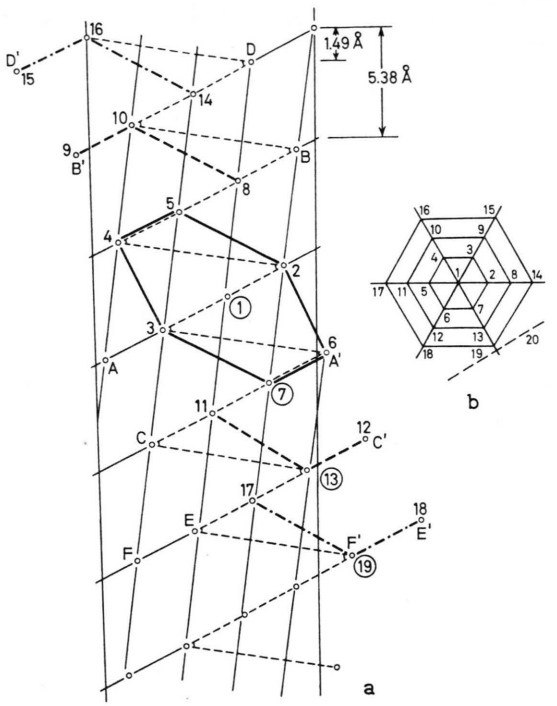

Fig. 2. Imaginary cylindrical surface of the α-helix as split open parallel to the axis of the cylinder and laid flat. Positions where residues are attached to the backbone are indicated as small circles, labeled A, B, ... or 1, 2, ... Dotted line (...) is the sketch of the backbone. The hexagon with residue positions 1 through 7 (solid line) and the additional layers, residues 8—10 and 11—13 (- - -) for the first augmented hexagon, also the residues 14—16 and 17—19 (— · — · —) for the second augmented hexagon are shown. When the cylinder is formed, D′, C′, ... coincide with D, C, ... Fig. b illustrates the three concentric hexagons covering corners 7, 13 and 19, all prime numbers less than 20.

which bring the number of neighboring corners to 13 and 19, respectively, Fig. 2 b. All three numbers 7, 13, and 19 are the prime numbers less than 20, just as 3 is less than 4.

It is, therefore, Pauling's α-helix that not only creates a surface but also guarantees a unique neighborhood pattern for each amino acid. Referring to Fig. 2, it should be noted that the primary structure of the chain on this α-helix surface is $(16-15-14) - (10-9-8) - (5-4-2) - 1 - (3-6-7) - (11-12-13) - (17-18-19)\ldots$.

Returning to the 7-member hexagon, the basic unit of a peptide chain on the α-helix, we observe the following. Just as in the case of DNA bases, forming triplets thus creating $4^3 = 64$ possibilities, for 20 amino acids forming hexagons creates $20^7 = 1.28 \times 10^9$ possibilities. On each hexagonal unit of the α-helix, one can then store one of the possible 1.28×10^9 *words*. It should be expected that not

only is this number more than enough as a meaningful size for the biological dictionary, but also degeneracy probably exists just as in the case of the genetic code, $64 \rightarrow (20+3)$. Therefore, the second and third layers leading to 13 and 19 neighborhood amino acids, respectively, may be meaningful in the grammar definition of the biological communication rather than replacing the basic unit of the hexagon.

The α-helix can be completely covered by non-overlapping hexagonal neigborhoods, Fig. 3. It is clear that on the cylindrical surface of the α-helix these hexagons are placed as not flat but bent surfaces; thus for example, the hexagon with corner A′ is the one with A, in opened flat form. The fictitious slitting of the α-helix done earlier, Fig. 2, is of course, an artificial one for visualizing the surface of the cylinder, but on the intact cylinder the hexagons are bent. The locus of centers of the hexagons follows a wavy line in the opened flat picture, **Fig. 3.**

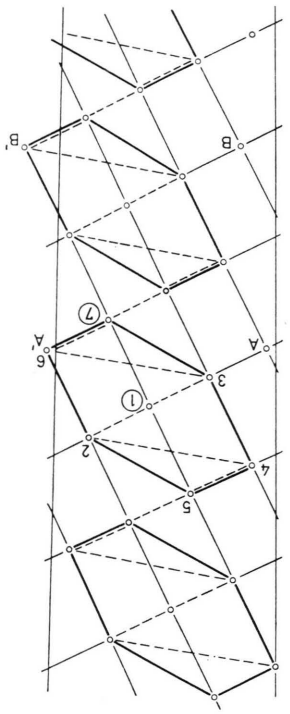

Fig. 3. An open helix showing a string of successive basic units, hexagons, (solid line) covering the entire residue positions on the helix. The backbone is sketched as (- - -).

Rather curious, but possibly significant, features of this helix in addition to those listed above are:

1. The basic pattern, the 7-member hexagon, is formed by triangulation, the triangle being the simplex of the two-dimensional space.

2. This leads also to the fact that, the third amino acid (Fig. 3, position 2) is also the nearest neighbor of the first one on the chain (position 5), because the peptide chain is placed on the α-helix surface.

3. Between every 5th amino acid location, there is a hydrogen bond.

4. The 3rd (position 2) and 5th (position 3) on the chain are nearest neighbors of the central (position 1) amino acid.

5. The number of amino acids including the first (1) on the hexagon and the central amino acid of the next hexagon is 11.

6. Counting all the atoms lying on a linear chain, from the H atom, say bonded to the N atom just after the carbon at position 5, to the 0 atom bonded to C just before the carbon at position 3, Fig. 3, the hydrogen bond between these H and O links the 13th atom to the first one on the helix.

It should be noted that the prime numbers 3, 5, 11, 13, 17, and 19, all less than 20, are meaningful in connection with the structure of the α-helix except 17.

Therefore, the α-helix is a basic structure of nature combining the mathematics (logic) of geometry with the chemistry (matter) to define the communication (biological activity). In order to eliminate misinterpretation of the present thesis, the conclusions which are drawn should be pointed out:

1. The α-helix provides the realization of a necessary set of words for the sentence to be expressed by each protein macromolecule.

2. The dictionary of words can then be enumerated by studying the α-helix regions of known and to-be-discovered protein molecules.

3. The possible grammar rules, Fig. 1, will become clear as the biological activity can be related to the physiological functions of each protein molecule.

4. All other existing theoretical and experimental studies on protein sequence may also be used as "data" for future studies based on the word of life, the hexagon.

5. While non-helical regions including β-sheets of molecules play various roles, some may well be significant also as "punctuations" and other syntax features of the biological language.

[1] T. T. Wu and E. A. Kabat,
a) Proc. Nat. Acad. Sc. U.S. **68**, 1501–1506 [1971];
b) Proc. Nat. Acad. Sc. U.S. **69**, 960–964 [1972];
c) Proc. Nat. Acad. Sc. U.S. **70**, 1473–1477 [1973];
d) J. Mol. Biol. **75**, 13–31 [1973];
e) Biopolymers **12**, 751–774 [1973].

[2] S. Erhan and L. D. Greller,
a) Int. J. Peptide Protein Res. **6**, 165–173 [1974];
b) Int. J. Peptide Protein Res. **6**, 175–181 [1974];
c) Nature **251**, 353–355 [1974].
[3] V. J. Hruby, Peptides and Proteins, An Survey of Recent Developments, **Vol.3**, pp. 1–188 (ed. B. Weinstein), Marcel Dekker, Inc., New York 1974.
[4] L. Pauling and R. B. Corey, Proc. Nat. Acad. Sc. U.S. **37**, 235–240 [1951].
L. Pauling, The Chemical Bond, p. 230, Cornell University Press, Ithaca, N. Y. 1967.