J. theor. Biol. (2001) **209,** 303–317
doi:10.1006/jtbi.2001.2265, available online at http://www.idealibrary.com on IDEAL®

Topological Nature of the Genetic Code

V. A. KARASEV*† AND V. E. STEFANOV‡

*St. Petersburg State Electrotechnical University, Prof. Popov str. 5, 197376 St. Petersburg, Russia and
‡Department of Biochemistry, St. Petersburg State University, Universitetskaya nab. 7/9,
199034 St. Petersburg, Russia

(*Received on* 7 *July* 2000, *Accepted in revised form on* 12 *January* 2001)

A model for topological coding of proteins is proposed. The model is based on the capacity of hydrogen bonds (property of connectivity) to fix conformations of protein molecules. The protein chain is modeled by an *n*-arc graph with the following elements: vertices (α-carbon atoms), structural edges (peptide bonds) and connectivity edges (virtual edges connecting non-adjacent atoms). It was shown that 64 conformations of the 4-arc graph can be described in the binary system by matrices of six variables which form a supermatrix containing four blocks. On the basis of correspondences between the pairs of variables in matrices and four letters of the genetic code matrices and supermatrix are converted, respectively, into the triplets and the table of the genetic code. An algorithm admitting computer programming is proposed for coding the *n*-arc graph and protein chain. Connectivity operators (polar amino acids) are assigned to blocks of triplets coding for cyclic conformations (G, A—in the second position), while anti-connectivity operators (non-polar amino acids) correspond to blocks of triplets coding for open conformations (C, U—in the second position). Amino acids coded by triplets differing by the first base have different structures. The third base for C, U and G, A is degenerated. Properties of the real genetic code are in full agreement with the model. The model provides an insight into the topological nature of the genetic code and can be used for development of algorithms for the prediction of the protein structure.

1. Introduction

The problem connected with the nature of the genetic code emerged when the assignment of coding triplets to encoded amino acids was represented by the table of the genetic code (Fig. 1) (Ycas, 1969). Pelc (1965) analysed the correlation existing between coding triplets and the structure of amino acids. It was noted that columns containing U, C (the second letter in the triplet) correspond mainly to hydrophobic (Phe, Leu, Ile, Val, Ala, Pro) and weakly polar (Ser, Thr) amino acids, whereas those containing A, G involve strongly polar amino acids (Asp, Glu, Asn, Gln, His, Arg, Lys). The third base does not influence significantly the coded residue (degeneration of the third base in triplets). Changes in one base of a triplet often correlate with minimal structural rearrangements in the side chains of amino acids.

The nature of "triplet–amino acid assignment" was first discussed by Crick (1968), who suggested that the modern code could emerge as a result of a "frozen accident" conserved in evolution. The idea was, however, criticized because of a number of deviations from the universal genetic code (Knight *et al.*, 1999). Certain progress has

† Author to whom correspondence should be addressed. E-mail: cmid@eltech.ru

0022–5193/01/070303 + 15 $35.00/0

1 \ 2	U	C	A	G	3
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STP	STP	A
	Leu	Ser	STP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Met	Thr	Lys	Arg	A
	Ile	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

FIG. 1. Table of the genetic code.

been recently achieved in the investigation of the spatial structure of the triplet code (Klump, 1993; Jimenez-Montaño *et al.*, 1996; Karasev & Sorokin, 1997). Klump (1993) proposed a concept of sequential spaces. Thus, purines and pyrimidines are represented by one symbol each: R for purines (A or G) and Y for pyrimidines (C, U or T). The spatial representation of such a two-letter code is a three-dimensional cube (first space). Each corner of this cube corresponds to a particular sequence (in R/Y notation) and all eight corners represent eight possible permutations (RRR, RRY, etc.). If bases are given their correct identities (binary for each position), then every corner of the basic cube contains another cube (subcube–second space). The overall dimension is now six and 4^3 codons can be represented in eight subgroups.

Jimenez-Montaño *et al.* (1996) suggested binary interpretation of the genetic code (Gray code). According to the authors, a great number of different Gray Codes can be associated with the Genetic Code depending on the order of importance of the bits in a code word. To justify their choice the authors analysed the significance of the chemical type of the bases (R–A, G–purines; Y–C, U–pyrimidines) and of the H-bonding character: weak (A, U) and strong (G, C) with two and three hydrogen bonds in Watson–Crick pairs, respectively. They assumed that the first bit in the binary coding is of chemical and the second one is of H-bonding character, and assigned the following correspondences $A = 00$, $G = 01$, $U = 10$, $C = 11$. The binary code of six variables can be represented by the Boolean hypercube B^6. The authors transformed the hypercube into the triplet genetic code. Vertices of the hypercube were labeled with the corresponding codons and one-letter symbols of amino acids were assigned to triplets. The authors (Jimenez-Montaño *et al.*, 1996) underline the significance of single base changes in the hypercube.

A principal novelty of the third variant of genetic code in the form of Boolean hypercube B^6 is that it includes the idea of the vicinity of the point (Karasev & Sorokin, 1997). The hypercube (Fig. 2) encloses more information than that described earlier (Yablonskii, 1986). Both triplets and encoded amino acids are placed in its 64 vertices. Triplets are related via single transitions of bases, so that neighboring triplets differ from one another by one base. The vertices form seven layers. The structure of the code is hierarchic. The elements are arranged in two groups. Eight quartets in the upper part of the structure (heavy line) code for one amino acid each (degeneration of the third base in triplets) and the other eight quartets code for two amino acids each. Each triplet from the upper group is related to a triplet from the lower group via Rumer transformation (Rumer, 1968): $C \leftrightarrow A$, $G \leftrightarrow U$, obeying C_2 symmetry. Thus, the vertices of the hypercube can be brought into one-to-one correspondence with the 64 six-digit Boolean variables, which, in turn, can be transformed into triplets of the genetic code.

Recently (Karasev, 1998; Karasev & Luchinin, 1998; Karasev *et al.*, 2000), a model for topological coding of chain polymers was proposed. The present paper focuses on application of the model to proteins and genetic code. On this basis isomorphism of the genetic code to Boolean hypercube B^6, the nature of "triplet-amino acid assignment" and characteristics of the code are analysed.

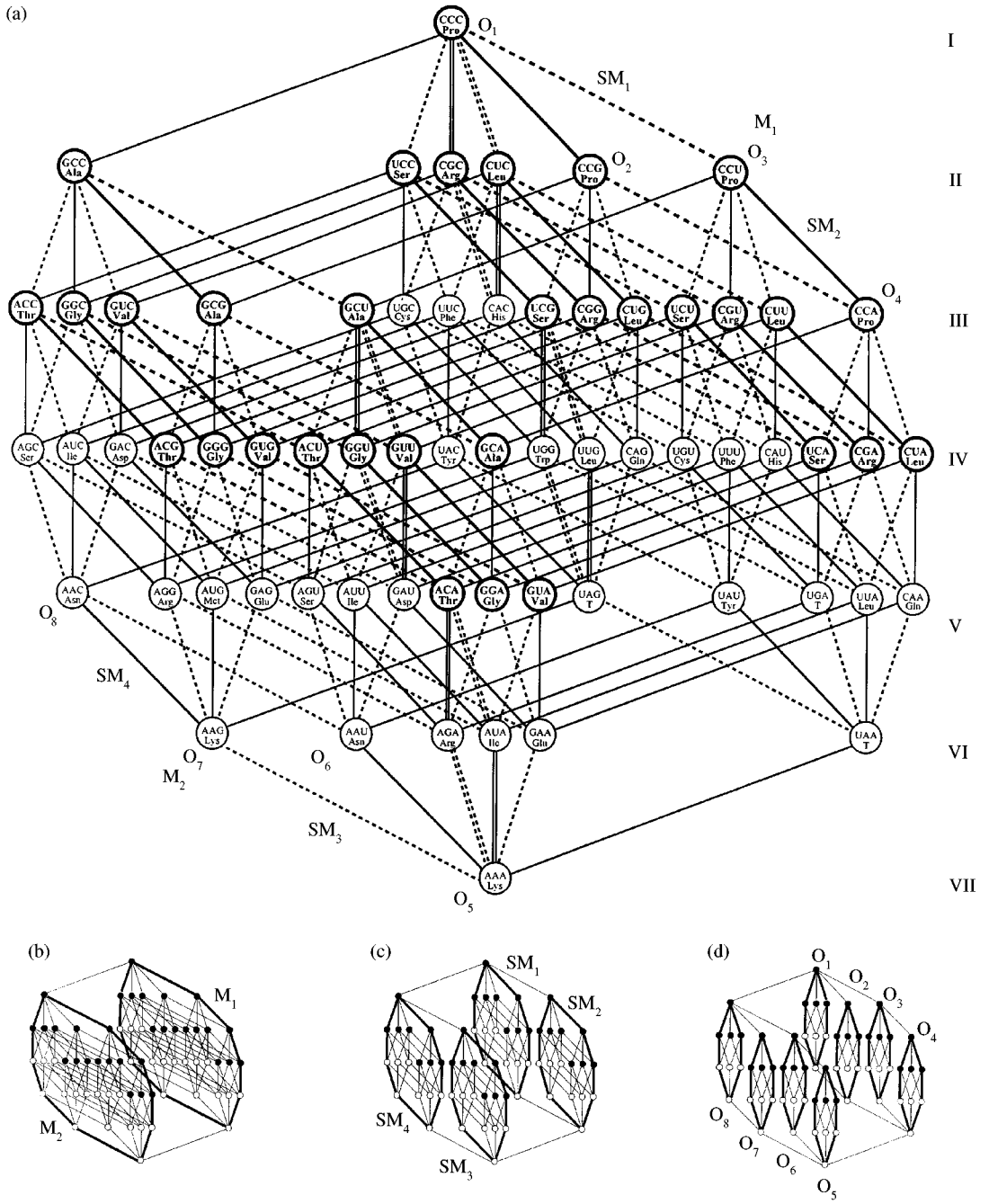

FIG. 2. Topological structure of the triplet genetic code. (a) general structure, isomorphic to Boolean hypercube B^6; (b)–(d) hierarchy of the hypercube structures: (b) two sets M_1 and M_2 comprising 32 elements each; (c) four sets SM_1–SM_4 comprising 16 elements each; (d) eight octets O_1–O_8; I–VII–strata.

2. The Model for Topological Coding of Proteins

2.1. DEFINITIONS

Earlier, the capacity of the main and side chains of chained polymers to fix the conformation of the latter was assumed as a prerequisite of their self-assembly (Karasev *et al.*, 2000). Such a capacity was called connectivity. Polypeptides are chained polymers possessing connectivity. Their conformation is fixed due to hydrogen bonds and other interactions.

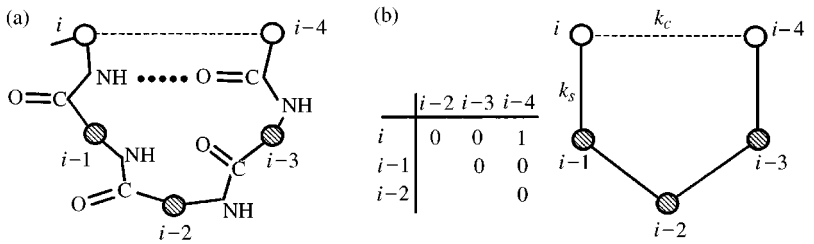

FIG. 3. Fragment of a linear chain polymer, its graph and matrix (b). (a) HN–C=O–groups of atoms connecting i–$i-1$, $i-1$–$i-2$, $i-2$–$i-3$ and $i-3$–$i-4$ atoms into the chain; atoms NH and O are connected by hydrogen bonds providing fixation of the ith and $(i-4)$-th atoms; (b) structural edges bind vertices i–$i-1$, $i-1$–$i-2$, $i-2$, $i-2$–$i-3$ and $i-3$–$i-4$; the connectivity edge connects vertices i–$i-4$; in the matrix it corresponds to 1: k_s—structural edge constant; k_c—connectivity constant.

The polypeptide chain can be represented by the n-arc graph (Karasev, 1998). The 4-arc graph [Fig. 3(b)] is a minimal model. Vertices $(i, i-1, \ldots, i-4)$ correspond to the α-carbon atoms of the periodically repeated unit (residues) of the protein molecule [Fig. 3(a)]. Structural edges k_s (solid lines), connecting vicinal vertices, represent corresponding peptide bonds. To model fragments of the protein molecule fixed due to hydrogen bonds, steric hindrance or otherwise, we close the graph with a "connectivity" (virtual) edge [Fig. 3(b), dotted line]. The connectivity edge is ascribed the value k_c, varying in the range 0–$2k_s$. In the latter case, the graph becomes linear chain without connectivity edges.

Let us denote the occurrence of a connectivity edge as "1" and the absence of such an edge as "0". Matrix representation of the connectivity states of the graph [Fig. 3(b)] is given in the same figure. In Fig. 3(a), the hydrogen bond of two peptide groups fixes i–$i-4$ α-carbon atoms, thus rendering them connected. In the matrix, this pair is represented by 1, while disconnected pairs of vertices are represented by 0. Protein conformations, connectivity states of the graph and their algebraic description were provided earlier (Karasev *et al.*, 2000). The general form of the matrix, describing the connectivity state of the 4-arc graph is shown in Scheme (1):

	$i-2$	$i-3$	$i-4$
i	x_1	x_2	x_3
$i-1$		x_4	x_5
$i-2$			x_6

(1)

i, $i-1, \ldots, i-4$ are vertices of the graph, $x_1, x_2, \ldots, x_6$ are variables assuming values 0 or 1. We shall also use a compact notation: $x_1x_2x_3x_4x_5x_6$.

2.2. SUPERMATRIX OF THE CONNECTIVITY STATES FOR THE 4-ARC GRAPH

All possible connectivity states of the 4-arc graph from completely extended (matrix containing only 0-elements) to completely connected (matrix containing six elements equal to 1) were considered. There are 64 connectivity states of the 4-arc graph and corresponding matrices (Fig. 4) They can be converged to a "supermatrix of connectivity states" (SCS). The SCS is constructed according to the following rules: rows are generated by the first pair of variables (x_1x_2) in the sequence 00, 10, 01, 11 and columns—by the third pair of variables (x_5x_6) in the sequence 00, 01, 10, 11. Thus, the SCS consists of four blocks, comprising 16 states each. The main property of each block is the occurrence of common second pairs of variables (x_3, x_4 — 00, 01, 10, 11).

One can find two types of symmetry in the SCS. One of them exists within blocks. Thus, in the first block ($x_3x_4 = 00$) the connectivity states of the first and third pairs are symmetrical, i.e. $00 \leftrightarrow 00$, $10 \leftrightarrow 01$, etc. The corresponding matrices and graphs are arranged symmetrically with respect to the main diagonal. A particular case of this symmetry is the intrinsic symmetry of matrices lying on the main diagonals, e.g. 000000, 100001, 010010, 110011.

The second type of symmetry is related to the structure of the SCS as a whole. Two groups of

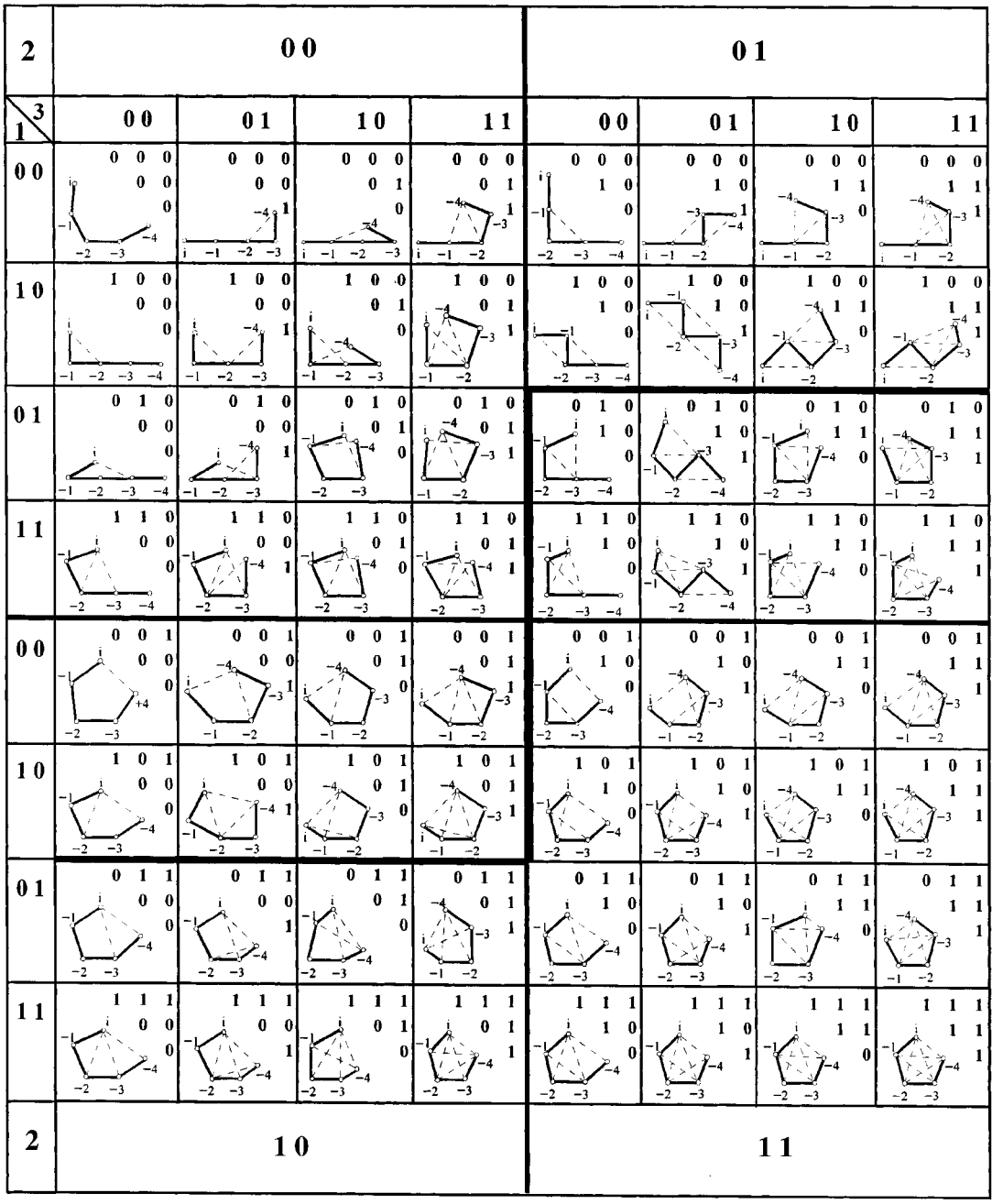

FIG. 4. Supermatrix of connectivity states for the 4-arc graph. Pairs of variables are marked with numbers in the left upper and lower corners of the table. Graphs are constructed for $k_c = \sqrt{2}$. Two groups of elements related by symmetry C_2 and antisymmetric transformation of variables ($0 \leftrightarrow 1$) are separated by a thick line.

matrices, in which 0-elements of the matrix belonging to one group correspond to 1-elements of the matrix of the other group and vice versa, e.g. 000000↔111111, occupy positions which are related by C_2 symmetry (separated by a solid line in Fig. 4). This type of symmetry was called antisymmetry (Karasev *et al.*, 2000).

From Fig. 4 it is seen that the connectivity pattern of the graphs is mainly determined by connectivity in $i - i - 4$ and $i - 1 - i - 3$ positions (pair of variables x_3x_4). Thus, in the block with $x_3x_4 = 00$ weakly connected graphs dominate (the number of the degrees of freedom 1 and 2). One can conceive it pulling the graph by one

of its vertices. If this does not necessarily cause displacement of two other vertices, the fragment is said to have a degree of freedom. Only the last graph of that block (variables 110011) has a stable connected conformation (no degree of freedom). In the block with $x_3x_4 = 01$, mainly graphs are present, having conformation with 2 to 1 degrees of freedom. The graph characteristic of this group is 100101. It corresponds to the β-sheet structure of protein. As in the previous block, the conformation of the last graph (110111) is the only stable one. It represents 3_{10} helix (Karasev, 1998). Blocks 00 and 01 contain mainly open conformations of the 4-arc graph.

In the next block ($x_3x_4 = 10$), due to connectivity between vertices i and $i-4$, the graph closes, forming a ring. Graphs in the upper left corner have unstable conformations inside the cycles (degrees of freedom from 2 to 1), while three graphs in the lower right corner have stable conformations (111010, 011011 and 111011)—equivalent of α-helix conformation. Finally, in the block with variables $x_3x_4 = 11$, there are eight stable conformations. They occupy the last three columns and the lower three rows, with the exception of element 011110. All of them represent α-helix (Karasev *et al*., 2000). Thus, blocks 10 and 11 contain only cyclic conformations. Spatial representation of the resulting supermatrix composed of "six variables" elements is Boolean hypercube B^6 (Karasev *et al*., 2000).

2.3. TRANSFORMATION OF SUPERMATRIX OF CONNECTIVE STATES INTO THE TRIPLET GENETIC CODE

Information about the graph structure presented in the matrix form cannot be used for transmission, reproduction and copying. It should be transformed into a suitable form of unbranched chain. Since the number of variables in the matrices describing connectivity states of the 4-arc graph, is equal to 6, i.e. three pairs, and the total number of matrices in the SCS is 64, coding becomes possible on the basis of the four-letter alphabet. For pairs of variables x_ix_{i+1} we introduce the notation XYZ:

$$x_1x_2 = X, \quad x_3x_4 = Y, \quad x_5x_6 = Z. \qquad (2)$$

The values 00, 10, 01, 11, assumed by x_ix_{i+1}, can be denoted by symbols C, U, G, A of the genetic code:

$$\mathrm{C} = 00, \quad \mathrm{U} = 01, \quad \mathrm{G} = 10, \quad \mathrm{A} = 11. \qquad (3)$$

Using this correspondence, we transform the SCS (matrices) into the code, shown in Fig. 5. Triplets appear together with the amino acids they code for, as in Fig. 1. Information on the structure of the 4-arc graph in terms of the four-letter code assumes the form of a linear chain.

The second letter of the triplet (Y), the same for the whole block, codes for variables x_3 and x_4 and contains the main information on the graph structure. Symmetric matrices describing symmetric conformations of the 4-arc graph are encoded by triplets arranged symmetrically with respect to the main diagonals of the blocks (Fig. 5). Antisymmetric matrices, related by symmetry C_2, are encoded by triplets, which transform into each other according to Rumer's rule: $\mathrm{C} \leftrightarrow \mathrm{A}$, $\mathrm{G} \leftrightarrow \mathrm{U}$ (compare Figs 5 and 2). Detailed analysis of the structure of the SCS matrix and its transformation into the table of the genetic code showed that 64 connectivity states of 4-arc graphs can serve as topological basis for the genetic code (Karasev *et al*., 2000).

2.4. ALGORITHM OF CODING FOR THE *n*-ARC GRAPH AND *n*-UNIT PROTEINS

Earlier (Karasev, 1998; Karasev & Luchinin, 1998; Karasev *et al*., 2000) we developed an algorithm* for coding n-arc graphs. In the present work we have adapted the method for the analysis of protein molecules. We assume that protein self-assembly proceeds co-translationally, i.e. synchronized with its synthesis (Ellis & Hartl, 1999), passing consecutively the stages from the short-range structure up to folding of the completed blocks into tertiary structure. The algorithm deals with the short-range structures. The short-range structure of the n-arc graph is described with the quasi-diagonal matrix (Q-matrix).

*A program was written on the basis of the algorithm (Karasev & Demchenko, 1998). Readers interested in the computer program may contact the authors.

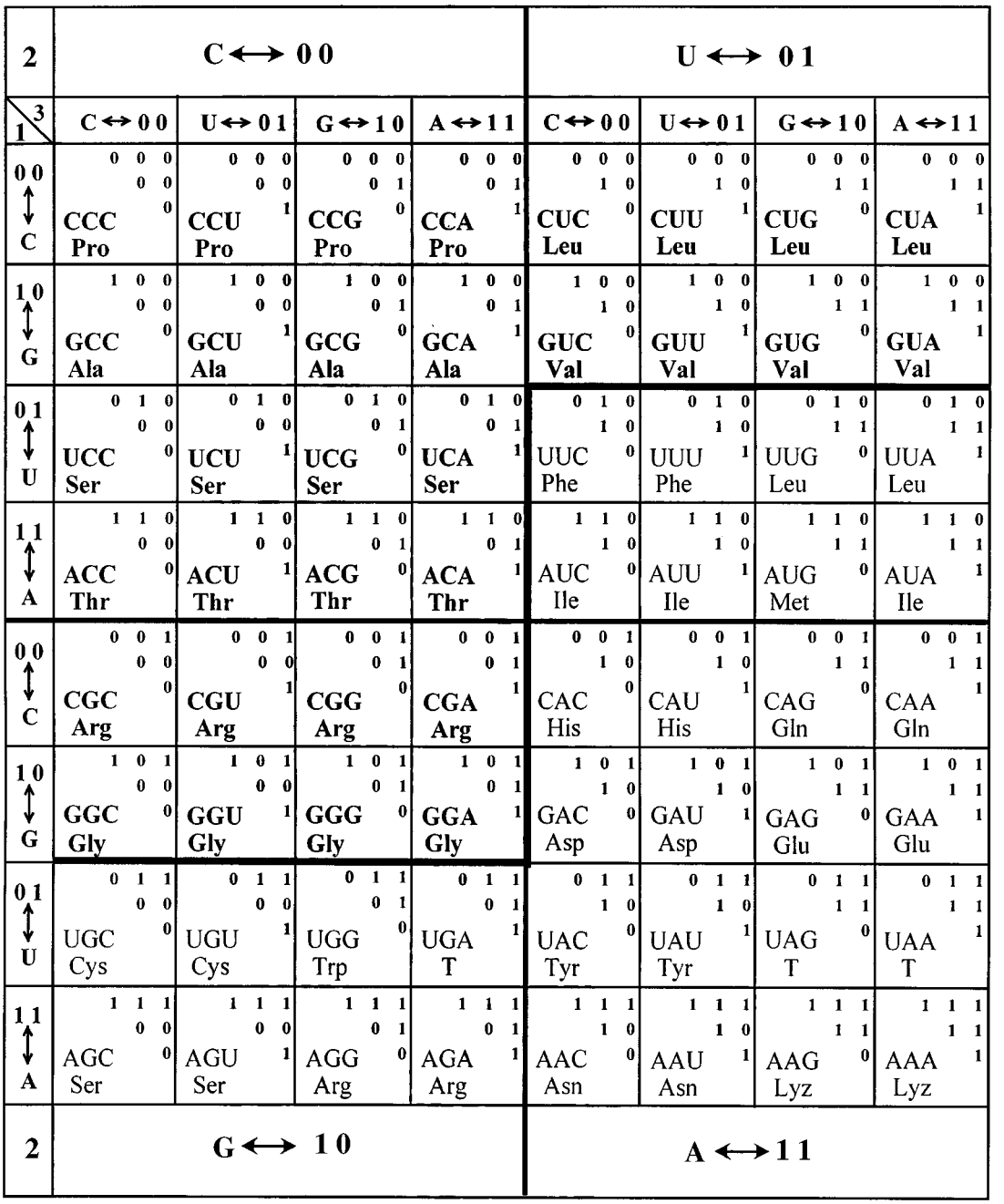

FIG. 5. Transformation of the supermatrix of connectivity states of the 4-arc graph into the genetic code.

At the initial stage of coding, matrices of six elements (marked with triangles in Fig. 6) are identified in Q-matrix, whose variables y_{ij}, assume values 0 or 1, similar to x_i. Variables y_{02}, y_{-11}, as well as y_{11}, y_{12} appear in two matrices, whereas y_{01}—in three. To encode the fragment each variable is expanded into components (arrows originating in Q-matrix, Fig 6). Connectivity assumes values 0 or 1, admitting disjunction formalism of summation [Scheme (4)]:

$$0 + 0 = 0, \quad 1 + 0 = 1, \quad 0 + 1 = 1, \quad 1 + 1 = 1. \tag{4}$$

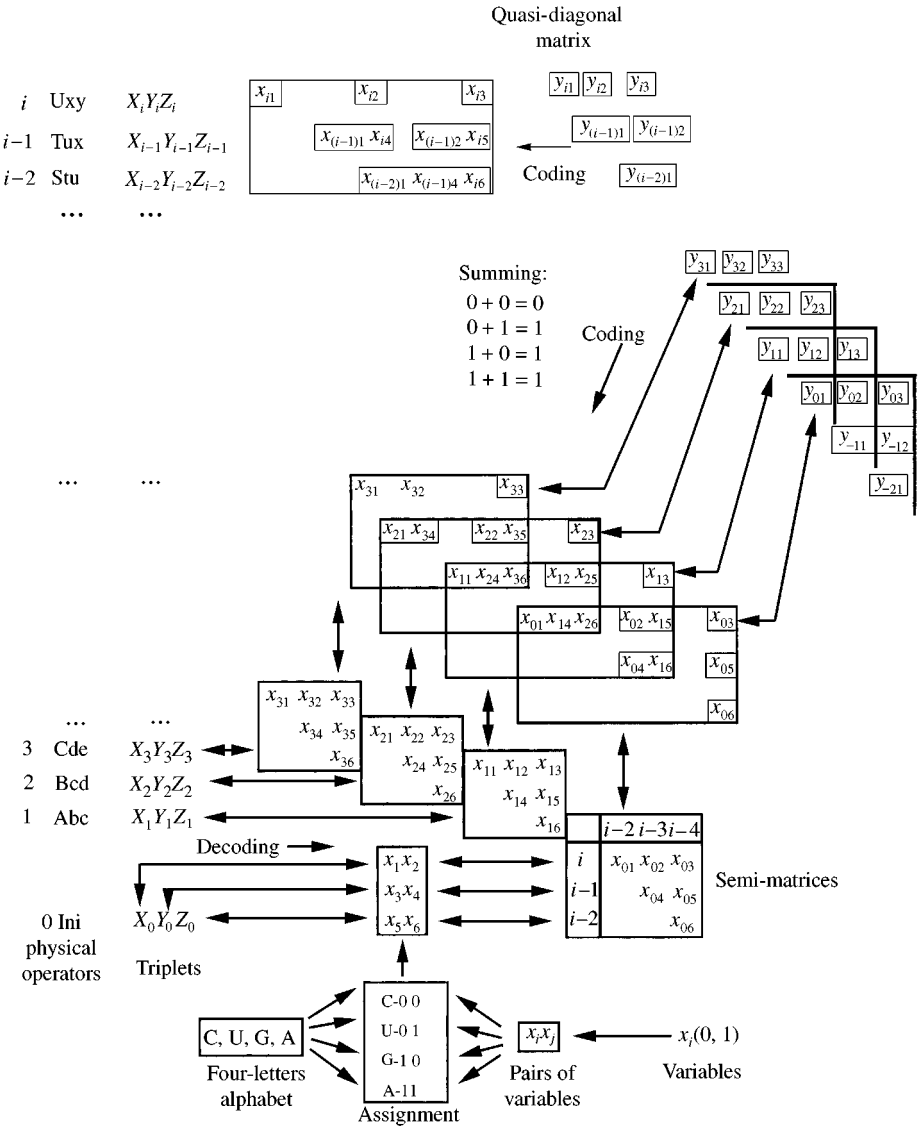

FIG. 6. Algorithm for coding—decoding of proteins.

Pairs of variables in the marked matrices are encoded as triplets and written down in order, as shown in Fig. 6: $X_0Y_0Z_0$, $X_1Y_1Z_1$, $X_2Y_2Z_2, \ldots, X_iY_iZ_i$. To reproduce the encoded structure, physical operators written down in the left column: Ini, Abc, …, Uxy (amino acids in the case of protein) should correspond to each triplet.

The decoding procedure is performed in the opposite direction. Triplets are decoded into matrices. Summation of variables follows the "disjunction" principle and results in the original Q-matrix. The algorithm implies an unambiguous decoding, whereas the same Q-matrix can be expanded by a number of ways. Therefore, one optimal graph topology can be used in coding proteins with different properties. Relevant examples have been reported in Ptitsyn & Finkelstein (1980).

3. Reconstitution of Proteins Structure and the Genetic Code

3.1. ASSIGNMENT OF PHYSICAL OPERATORS TO TRIPLETS

Two specific features of conformations of the 4-arc graph within SCS have been mentioned above: occurrence of two types of conformation—open (blocks 00 and 01) and close helical

(blocks 10 and 11); symmetry of pairs of conformations within each block. Physical operators should reflect these features. The most stable conformation of the protein chain is α-helix (Lim & Aglyamova, 1999). A newly synthesized fragment will tend to assume the optimal, i.e. helical conformation. Such a fragment bonded on the ribosome to tRNA and the corresponding matrix is shown in Fig. 7. The only site that can be affected by the side chain R of the just bound *i*-th amino acid is the area where the hydrogen bond between groups O=C–N_iH and O_{i-4}=C–NH is formed. This is represented by variable x_3 in the matrix. Connectivity and anti-connectivity operators can be distinguished by their mode of action.

3.1.1. *Connectivity Operators*

Connectivity operators are amino acid side chains which provide additional fixation of a four-element fragment, e.g. due to hydrogen bonds, in accordance with the encoded fragment of the 4-arc-graph. The generalized form of the connectivity operator is shown in Fig. 7(b). The terminal group Q=R–XH of the side chain of the *i*-th amino acid is capable of forming a hydrogen bond with group O_{i-4}=C–NH (Fig. 7), contributing the variable $x_3 = 1$ into the matrix.

3.1.2. *Anti-connectivity Operators*

Anti-connectivity operators are amino acid side chains which obstruct formation of a four-element fragment in accordance with the encoded fragment of the arc-graph. The side chain of the anti-connectivity operator, shown in Fig. 7(c), builds in the zone of the hydrogen bond of the main chain and thus obstructs formation of the four-element cycle (variable $x_3 = 0$ in the matrix).

It follows from Fig. 7(b) and (c) that connectivity operators (polar amino acids, variable $x_3 = 1$) must be assigned to blocks 10 and 11, whereas anti-connectivity operators (non-polar amino acids, variable $x_3 = 0$)—to blocks 00 and 01. Thus, after transformation into the topological genetic code the connectivity operators will be assigned to the first two blocks of triplets (G = 10 and A = 11) and the anti-connectivity operators to the other two (00 = C and 01 = U). This is supported by the arrangement of polar and non-polar amino acids in the code (Fig. 5).

3.2. AMINO ACIDS AS PHYSICAL OPERATORS

To identify sites of amino acids, which may act as physical operators, we analysed 32 protein structures. We found cyclic fragments, where side chains of amino acids form hydrogen bonds with the main chain, and arranged them in the table

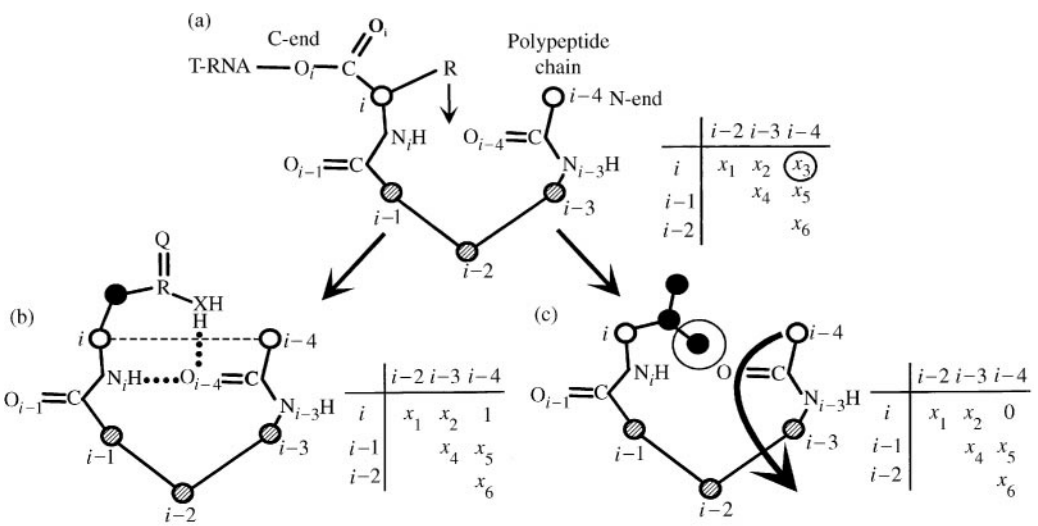

FIG. 7. Physical operators. (a) four-element protein fragment; (b) connectivity operator; (c) anti-connectivity operator. ○—electron shell radius of the methyl group. ●—carbon atoms.

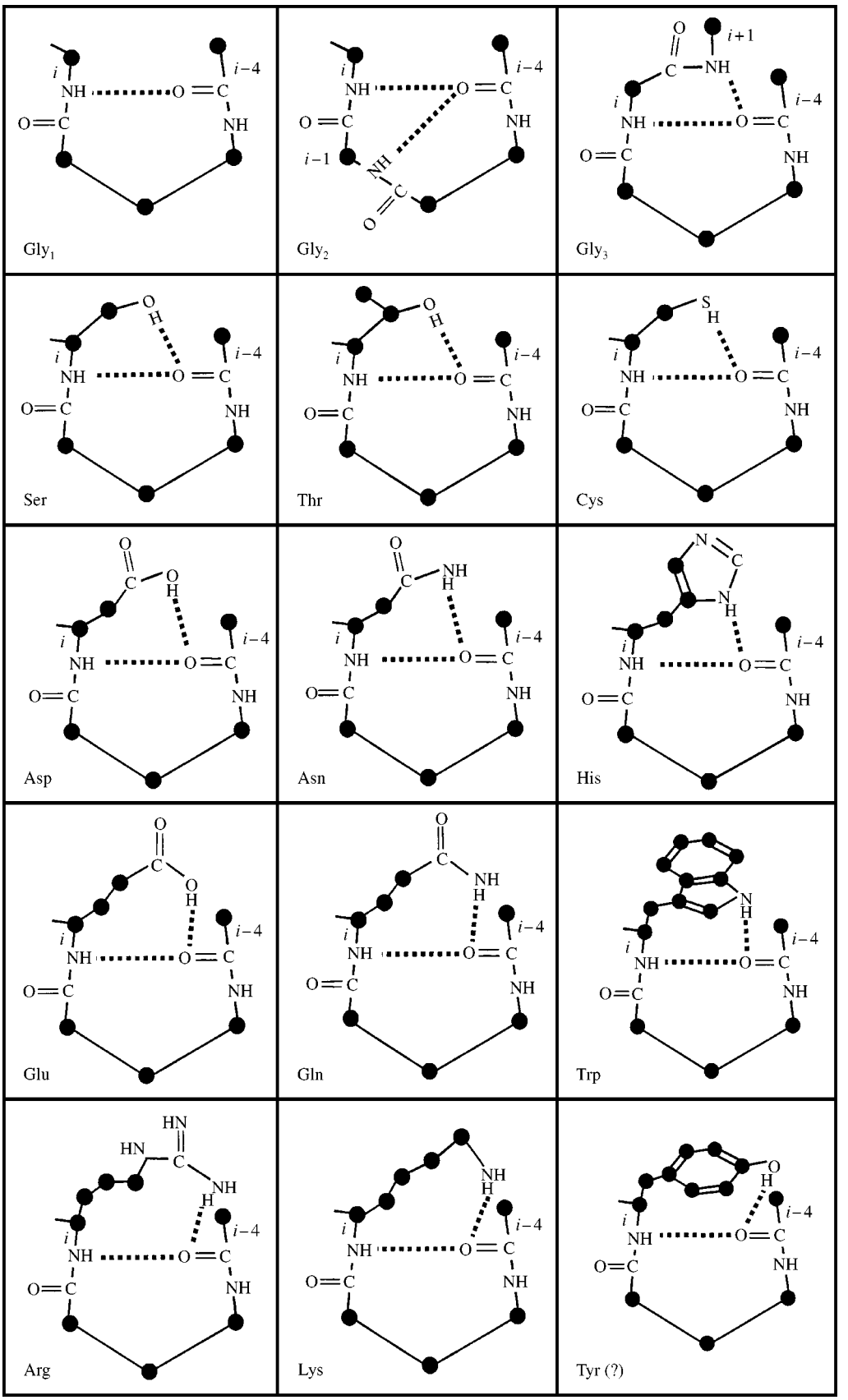

FIG. 8. Side chains of polar amino acids as connectivity operators.

(Fig. 8). Conformations of the amino acids are shown as they are encountered in proteins. The table begins with Gly_1, which has no side chain and can be considered as part of a polypeptide chain with the property of connectivity. Its HN–C=O-group between amino acids i and $i+1$ often forms an additional hydrogen bond with carbonyl O=C-group of amino acid $i-4$ (Gly_3). Practically all other amino acids containing polar groups in the side chain are capable of forming hydrogen bonds with the main chain thus providing fixation of a four-element fragment. The only exception is Tyr, whose cyclic fragment was not recorded in the chosen proteins. Hence, amino acids shown in Fig. 8 exhibit features of connectivity operators. It was also revealed that more than 90% of such bonds are formed between the side chain of the ith residue and the C=O-group of the $(i-4)$th residue.

Depending on the length of the side chains the slope of hydrogen bonds varies from left (for Gly_2, Ser, Thr and Cys) to right (for Arg and Lys) (Fig. 8). This indicates that the limiting size of the side chains of polar amino acids can be specified by their zone, where they function as physical operators.

In Fig. 9, side chains of non-polar amino acids are shown in the conformation of anti-connectivity operators. Pro is a typical anti-connectivity operator, which obstructs formation of the four-element cycle. N atom of this amino acid is located in the five-atom cycle and is not able to form a hydrogen bond with O_{i-4}=C–NH-group. It is just Pro that terminates the α-helix. It is logical to assign zero connectivity to Pro. Comparing Figs 4 and 5, one can see that all four triplets of the first row in the block C=00 code for conformations of the 4-arc graph with the minimal connectivity, and it is just these conformations that correspond to Pro.

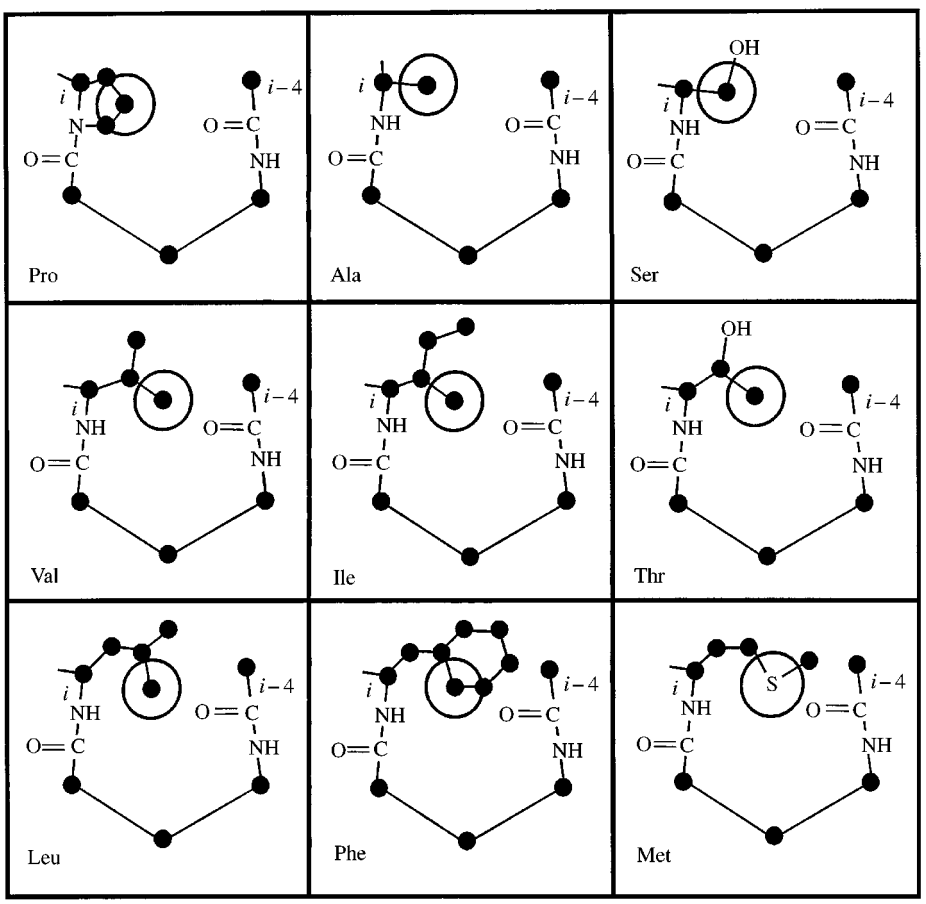

FIG. 9. Side chains of non-polar and weakly polar amino acids as plausible anti-connectivity operators.

Side chains of the amino acids Val, Ile and Thr (the group of Val) have two substituents in β-position, which also obstruct formation of the four-element close cycle. Val is the most efficient amino acid in forming β-structure (Schulz & Schirmer, 1979). Comparing Figs 4 and 5, we see that triplets, corresponding to Val, code for β-structure conformations of the 4-arc graph. The amino acids Leu and Phe (the group of Leu) have two substituents in γ-position. Their behavior must be similar to that of the former group.

We presume that Ser and Thr, manifesting a dual character, can function both as anti-connectivity and connectivity operators, which are present in both groups of operators (Figs 8 and 9). As in the case of connectivity operators, one can presume that the limiting size of the side chains of non-polar amino acids is also specified by their zone, where they function as operators of anti-connectivity.

3.3. RECONSTRUCTION OF SYMMETRIC CONFORMATIONS

To conceive physical operators which reconstruct symmetric conformations, one should take into account that in the process of biosynthesis amino acids bind to tRNA in the same position (Kaziro, 1978). Only side chains of amino acids change. Suppose that these are connectivity operators with similar properties but of different size (Fig. 10). Let us denote functional groups situated at the end of the chains as $Q_1{=}R{-}X_1$ and $Q_2{=}R{-}X_2$. As seen from Fig. 10, hydrogen bonds of two side chains of different lengths have different slope and, hence, differently directed field lines. Connectivity of the i-th–$(i-4)$-th α-carbon atoms in the two cycles is the same (dotted line), whereas connectivity of other atoms is different. For the left operator, the traction force is directed to the left (dotted line with arrows) so that the i-th–$(i-2)$-th atoms become connected. For the right operator this force is right directed and connected are atoms $(i-2)$-th–$(i-4)$-th. The above is reflected in matrices $101x_4x_5x_6$ and $x_1x_21x_401$. The two operators (side chains) differ by their length and generate symmetric conformations (Fig. 10). Apparently, this can be recognized as a general principle of triplet–amino acid assignment in the genetic code. The reasoning is, apparently, applicable to the anti-connectivity operators as well.

In fact, triplets arranged symmetrically with respect to the main diagonals of the blocks of the genetic code, code for non-identical amino acids (Fig. 5): Ala–Pro, Ser–Pro, Thr–Pro, Ser–Ala, Thr–Ala, Thr–Ser occupy symmetric positions in block C = 00; in block U = 01 these are Val–Leu, Phe–Leu, Phe–Val, Ile–Leu, Ile–Val, Met–Leu, etc. (Fig. 5). There are six Leu in that block. However, pairs of Leu do not occupy symmetric positions. Neither do six Arg of the block G = 10. Similarly, in the block A = 11, Asp and Asn, though similar in size, do not occupy symmetric positions, neither do Glu and Gln.

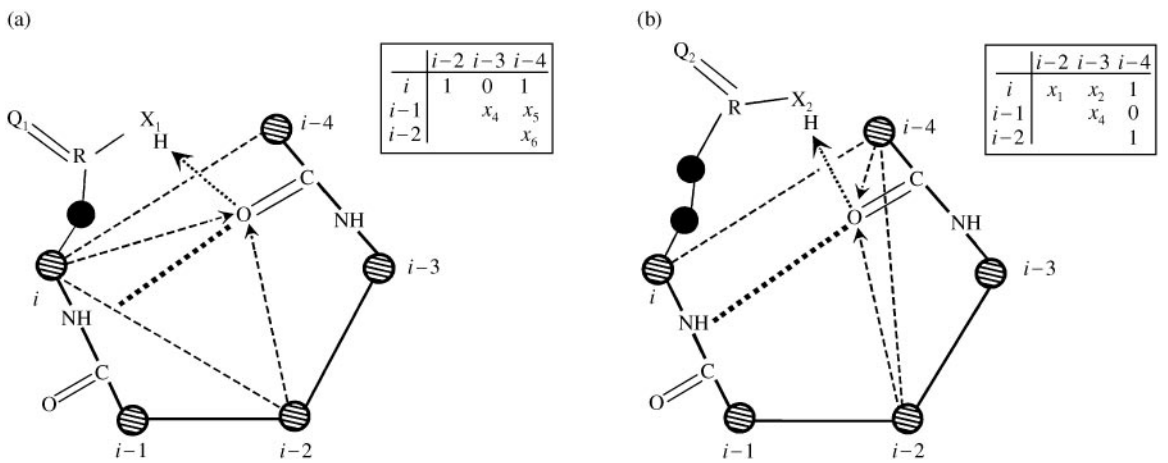

FIG. 10. Reconstitution of symmetrical conformation by different physical operators (side chains of polar amino acids of different lengths).

Thus, the triplet–amino acid assignment, observed in the genetic code does not disagree with the idea that symmetric conformations are reconstructed by different physical operators. There is no problem of reconstruction of symmetric conformations, situated on the main diagonals, owing to the degeneration of the third letter in the triplet (see Section 3.4).

It should be also noted that both polar and non-polar amino acids can function as physical operators only if their orientation is towards protein, as shown in Fig. 7. They act upon the already synthesized structure, i.e. they are retro-operators. That is why all amino acids must have the same stereo configuration.

3.4. CONSEQUENCES OF THE CODING ALGORITHM

Let us consider coding–decoding of the i-th triplet, shown in Fig. 6, which represents the i-th stage of protein synthesis [Scheme (5)]:

depends both on the values of x_{i4} and x_{i5} (0 or 1) and on the probability of the particular meaning assumed by the first letter in the $(i-1)$-th triplet ($X_{i-1} = x_{(i-1)1}x_{(i-1)2}$). The latter event is determined by evolutionary factors and requires a special investigation.

Of major interest is variable x_{i6}, which appears in the sum: $x_{(i-2)1} + x_{(i-1)4} + x_{i6}$. There are eight possible items: 000, 001, 010, 011, 100, 101, 110, 111, of which seven include value 1, so that addition carried out according to disjunction logic gives 1. Therefore, the probability of value "0" appearing is negligible (1/8). Hence, the value of variable x_{i6} (0 or 1) is no more of any importance. This means that the part of the 4-arc fragment of the protein structure described by variable $x_{i6}(i-2–i-4)$ is completely formed and cannot be influenced by the i-th physical operator. Therefore, the operators, which correspond to triplets with the third letter C = 00 and

No.	AA	Triplets	Matrices			Q-matrix			
i	Uxy	$X_i\ Y_i\ Z_i$	x_{i1}	x_{i2}	x_{i3}	y_{i1}	y_{i2}	y_{i3}	(5)
$i-1$	Tux	$X_{i-1}\ Y_{i-1}\ Z_{i-1}$		$x_{(i-1)1}x_{i4}$	$x_{(i-1)2}x_{i5}$		$y_{(i-1)1}$	$y_{(i-1)2}$	
$i-2$	Stu	$X_{i-2}\ Y_{i-2}\ Z_{i-2}$			$x_{(i-2)1}x_{(i-1)4}x_{i6}$			$y_{(i-2)1}$	

The procedure carried out according to Scheme (5) implies that variables x_{i1}, x_{i2}, x_{i3} are the most important in the specific activity of the physical operator. Variable x_{i3} is common for the whole block, while $x_{i1}x_{i2}$ correspond to the first letter of the triplet (X_i). Hence, this letter accounts for the difference between the side chains of operators placed in neighboring rows. In fact, from Fig. 5 one can see that neighboring rows are occupied by different amino acids, e.g. in the first block these are Pro ($x_{i1}x_{i2} = 00$), Ala (10), Ser (01), Thr (11) and in the second—Leu (00), Val (10), Phe–Leu (01), Ile–Met (11), etc.

That part of the graph conformation, which is coded for by the second letter, can be completely realized only for variable x_{i3} corresponding to the true value y_{i3}. The type of conformation (open or closed) therewith is reproduced by the physical operator.

At the $(i-1)$-th stage, variable x_{i4} is added to $x_{(i-1)1}$ and x_{i5} to – $x_{(i-1)2}$, whose performance

U = 01, as well as G = 10 and A = 11, should be the same.

The obtained consequences are well illustrated by the genetic code (Fig. 5). One can see the pairs of amino acids through all the structure of the genetic code, e.g. Phe–Phe (C, U), Leu–Leu (G, A) Ser–Ser (C, U), Arg–Arg (G, A), etc. In all blocks of the code, in half of the cases, one triplet in a pair, coding for the same amino acid, is from the main diagonal (Fig. 5). Therefore, no special operators are necessary for the reconstruction of symmetric conformations of the 4-arc graph. The fact that in the upper part of the code four triplets coincide (variable x_5 is of no importance) has no effect on our conclusion about variable x_6. The only exceptions are amino acids Met and Ile in the block U = 01 and Trp and stop-codon (T) in the block G = 10, which also often form pairs in the mitochondrial code (Knight *et al.*, 1999).

The analysis of amino acids as physical operators, carried out in this section, can be used in development of an algorithm for prediction of protein structures.

4. Conclusion

The proposed model explains the occurrence of 64 triplets in the code by the fact that the encoded object is the 4-arc graph (analog of the helical fragment of protein) which can assume 64 connected conformations. Their matrix description leads to the structure of the triplet genetic code isomorphic to Boolean hypercube B^6. Linear non-overlapping pattern of genetic messages results from the algorithm of coding developed for *n*-arc graph using the quasi-diagonal matrix. It was shown that the matrix can be converted into a linear chain of triplets. The model explains the nature of "triplet–amino acid assignment". The capacity of the polypeptide chain to fold spontaneously, accompanied by formation of continuous HN–C=O-groups, is regulated by the side chains of amino acids. The latter are regarded as connectivity (polar amino acids) and anti-connectivity (non-polar amino acids) physical operators. Polar amino acids must be ascribed to triplets with G, A in the second position (they code for cyclic conformations), whereas non-polar—to triplets with C, U in the second position (they code for weakly connected and open conformations). The algorithm accounts for the base degeneration in the third position. By and large, properties of the real genetic code are in full agreement with the model. Thus, from the viewpoint of the model, the triplet–amino acid assignment suggests reconstitution of protein conformations encoded by triplets rather than manifestation of a "frozen accident" (Crick, 1968).

The model of topological coding of chained polymers is, apparently, applicable not only to proteins but also to other classes of biological chained polymers, particularly, to nucleic acids. One can presume that nitrous bases can form a system of physical (connectivity and anti-connectivity) operators. This makes clear why t-RNA and ribozymes of complex topology contain a great number of methylated nitrous bases.

The proposed model has certain limitations. It does not deal with the functions which amino acids may have in the protein structure. It neither addresses nor answers the question why a particular set of amino acids is realized. It may be expedient to complement the approach with the concept (Karasev *et al*., 1994) treating polar amino acids as functional modules (elements of initiation, delay, inversion, etc.) of conjugated ion–hydrogen bonds systems.

In general, the proposed model provides a new insight into the topological nature of the genetic code and can be used for development of algorithms for the prediction of the protein structure.

We are grateful to V. V. Luchinin for his support of this work and useful discussion of the paper. We also highly appreciate facilities offered in the laboratory of Dr Madariaga from Barcelona University during preparation of the manuscript and interesting discussions of the material. The work was financially supported by the research grants from Russian Foundation for Basic Research (RFFI) code 99-04-49836 and GR/CMID No. 50 from Ministry of Education of Russian Federation.

REFERENCES

CRICK, F. H. C. (1968). The origin of the genetic code. *J. Mol. Biol.* **38,** 367–379.

ELLIS, R. J. & HARTL, F. U. (1999). Principles of protein folding in the cellular environment. *Curr. Opin. Struct. Biol.* **9,** 102–110.

JIMENEZ-MONTAÑO, M. A., DE LA MORA-BASACEZ, C. R. & PÖSCHEL, TH. (1996). The hypercube structure of the genetic code explains conservative and non-conservative aminoacid substitutions in vivo and in vitro. *BioSystems* **39,** 117–125.

JUKES, T. H. (1990). Genetic code 1990. Outlook. *Experientia* **46,** 1149–1157.

KARASEV, V. A. (1998). How the topology of a biochip can be coded? *Biotekhnologia* **3,** 62–74.

KARASEV, V. A. & DEMCHENKO, E. L. (1998). Computer program "Decoder of Proteins Supramolecular Structure—Protein 3D" Registered in Russia copyright agency for computer programs, data bases and topologies of micro-schemes, No. 980143 from 03.05.98.

KARASEV, V. A., DEMCHENKO, E. L. & STEFANOV, V. E. (2000). The topological coding of polymers and protein structure prediction. In: *Chemical Topology*: *Applications and Techniques* (Bonchev, D. & Rouvray, D. eds). *Ser. Math. Chem.*, Vol. 6, pp. 295–345. New York, London, Paris: Gordon & Breach.

KARASEV, V. A. & LUCHININ, V. V. (1998). The problems of construction artificial bionic micro- and nanosystems. *Izvestija Vuzov. Ser. Elektron.* **6,** 5–15 (in Russian).

KARASEV, V. A., LUCHININ, V. V. & STEFANOV, V. E. (1994). A model of molecular electronics based on the concept of

conjugated ionic-hydrogen bond systems. *Adv. Mater. Opt. Electron.* **4,** 203–218.

KARASEV, V. A. & SOROKIN, S. G. (1997). Topological structure of the genetic code. *Russian J. Genet.* **33,** 622–628.

KAZIRO, Y. (1978). The role of GTP in polypeptide chain elongation. *Biochim. Biophys. Acta* **505,** 95–127.

KLUMP, H. H. (1993). The physical basis of the genetic code: the choice between speed and precision. *Arch. Biochem. Biophys.* **301,** 207–209.

KNIGHT, R. D., FREELAND, S. J. & LANDWEBER, L. F. (1999). Selection, history and chemistry: three faces of the genetic code. *Trends Biochem. Sci.* **24,** 241–247.

LIM, V. I. & AGLYAMOVA, G. V. (1999). The principles of formation of spatial structures of proteins and nucleic acids. Stereochemical modeling. *Mol. Biol.* **33,** 1027–1034.

PELC, S. R. (1965). Correlation between coding-triplets and amino-acids. *Nature* **207,** 597–599.

PTITSYN, O. B. & FINKELSTEIN, A. (1980). Similarities of protein topologies: evolutionary divergense, functional convergence or principle of folding? *Quart. Rev. Biophys.* **13,** 339–386.

RUMER, YU. B. (1968). Systematization of codons in the genetic code. *Dokl. Acad. Nauk SSSR* **183,** 225–226 (in Russian).

SCHULZ, G. E. & SCHIRMER, R. H. (1979). *Principles of Protein Structure.* New York, Heidelberg, Berlin: Springer-Verlag.

YABLONSKII, S. V. (1986). *Basics of Discrete Mathematics.* Moskow: Nauka (in Russian).

YCAS, M. (1969). *The Biological Code*, 351pp. Amsterdam, London: North-Holland.