Series in Mathematical Biology and Medicine. Vol.8.

Proceedings of the International Conference. "Advances in Bioinformatics and its applications".

Nova Southeastern University, Fort Lauderdale, Florida, USA 16-19 December 2004.

Eds.: M. He, G. Narasimhan, S. Petoukhov

2005

World Scientific Publishing Co. Pte.Ltd..

New Jersey – London – Singapore – Berlin – Shanghai - Hong Kong – Taipei - Chennai.

A DODECAHEDRON-BASED MODEL OF SPATIAL REPRESENTATION OF THE CANONICAL SET OF AMINO ACIDS

V.A. KARASEV, V.V. LUCHININ

St.-Petersburg State Electrotechnical University,
Prof. Popov str. 5, St.-Petersburg, 197376 Russia
e-mail: cmid@eltech.ru

V.E. STEFANOV

Department of Biochemistry, St.Petersburg State University,
Universitetskaya nab. 7/9, St.-Petersburg, 199034 Russia

In order to get a new insight into the nature of the canonical set of twenty amino acids we undertook an attempt to build up a spatial structure arranging the set of amino acids, like it had been done for duplets and triplets of the genetic code. Analysis of the properties of a set of 12 meridian cycles obtained on the structure of the duplet genetic code, which is isomorphic to Boolean hypercube B^4, revealed four groups of cycles united in pairs by anti-symmetry transformations of two types. These transformations become most illustrative when shown on the icosahedron, a polyhedron with 12 vertices. Related to the icosahedron is another polyhedron – dodecahedron, which has 20 vertices. Approach based on the use of the two polyhedrons was applied to the analysis of structure of the canonical set of 20 amino acids. It was demonstrated that four groups of amino acids, each containing five amino acids connected by anti-symmetry transformations of two types, can be distinguished in the initial set. The revealed principles were pictorially represented on the structure of dodecahedron.

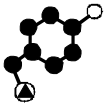
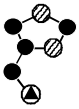
1. Introduction

Development of spatial models of the duplet and triplet genetic code isomorphic to Boolean hypercubes B^4 and B^6, respectively, is an important achievement [1-4]. However, the proposed structures deal with the duplet and triplet code only, ignoring the nature of the canonical set of amino acids. Only 20 amino acids are encoded by the genetic code and participate in creating protein molecules [5, 6]. This set must have its structural principles, which up to now remain obscure.

There are several approaches to classification of the amino acids' side chains. The most common one [6] implies classification of amino acids according to the physicochemical properties of their radicals (Table 1). One can mark two interesting features of the Table 1. First, most amino acid side chains form pairs, e.g. Leu – Ile, Ser – Thr, Asp – Glu, Arg – Lys, Asn – Gln, Phe – Tyr. To some extent, this is also valid for Cys and Met, His and Trp. Second, there are amino acids with opposite properties. Most distinctly this feature is

expressed in the group: Asp $^{-}$, Glu^{-} and Lys $^{+}$, Arg^{+}. Opposite properties can be revealed within other groups of amino acids as well. So far these grouping properties of the canonical set of amino acids have not been rationally explained [5, 6]. There is alternative approach based on the particular character of functioning of amino acids in the protein structure [7].

Table 1. Structure of side chains of canonical amino acids

Nonpolar and weakly polar			
 Proline (Pro)	 Glycine (Gly)	 Alanine (Ala)	 Valine (Val)
 Leucine (Leu)	 Isoleucine (Ile)	 Serine (Ser)(-)	 Threonine (Thr)(-)
Polar			
 Asparagic acid (Asp^{-})	 Glutamic acid (Glu^{-})	 Lysine (Lys^{+})	 Arginine (Arg^{+})
Neutral and Sulfur-containing			
 Asparagine (Asn)	 Glutamine (Gln)	 Cysteine (Cys^{-})	 Methionine (Met)
Cyclic			
 Phenylalanine (Phe)	 Tyrosine (Tyr^{-})	 Histidine (His^{+})	 Tryptophane (Trp)

Notation: black circles denote carbon atoms with different number of hydrogen atoms (CH, CH_2, CH_3); little empty circles – oxygen atoms (O, OH); big empty circles –sulphur atoms (S, SH); dashed circles – nitrogen atoms (N, NH, NH_2, NH_3); circles with enclosed black triangles – α-carbon atoms (CH).

The genetic code provides a natural basis for classification of amino acids. One of the classification based on the genetic code, addresses the idea of complementarity of amino acids encoded by complementary triplets [8-11]. The

authors [8, 9] undertook analysis of such amino acids and derived three groups of connectivity within them based on antiparallel complementarity of triplets.

Within the model developed by us [12] side chains of amino acids, encoded by triplets, are treated as physical operators reconstructing the encoded structure. They form two groups in the code table. The connectivity operators (polar amino acids) are encoded by codons with G and A occupying the second position (lower group), whereas in anti-connectivity operators (mainly non-polar amino acids) this position is occupied by C and U (upper group).

However, the system of 20 amino acids may have its own spatial representation only indirectly connected with the code. The present study aims at the analysis of the structure of the canonical set of amino acids, based on principles of symmetry and anti-symmetry, and development of a spatial model, which would provide an illustrative representation of the structure of the set. Earlier this problem was addressed in a preliminary study [13].

2. Construction of the model

2.1. *Prerequisites*

There are several prerequisites of the dodecahedron-based model of the structural arrangement of amino acids within the canonical set.

1. The number of amino acids in the canonical set coincides with that of the dodecahedron vertices – 20 [14].
2. Dodecahedron-icosahedron form dual system: the centers of twenty triangle facets of the icosahedron correspond to the vertices of the dodecahedron. Reciprocally, centers of twelve pentagon facets of the dodecahedron correspond to the icosahedron vertices [14].
3. The side chains of amino acids are involved in the process of spatial and temporal self-organization of the protein molecules [12]. This model agrees with the idea of the co-translational mechanism of protein molecule folding [15], which has been experimentally verified. In addition, on the basis of the analysis of triplet-amino acids correspondences it was shown that amino acid side chains with different structure are needed to generate symmetric conformation [12]. Thus, one can presume that the ideas of symmetry should lie in the topological basis of the genetic code.
4. Spatial structures isomorphic to Boolean hypercubes can be used for describing the processes of self-organization of protein molecules. A method of interpreting temporal processes based on the structures isomorphic to Boolean hypercubes was used for the analysis of the mathematical structure of the Ancient Chinese calendar [16]. Basically, the method suggests that displacement on the structure of the Boolean hypercube from the vertex with values of variables equal to 0 to the vertex

with the variables equal to 1 represents a model of discrete changes in time of the analyzed object. The resultant paths were called “meridian cycles” (M-cycles). The methodological connection between the amino acids and M-cycles, lies in the fact that, according to the definition, M-cycles are the elements of the temporal and spatial description of the structural changes of the 4-arc graph, whereas the amino acid are regarded as functional operators generating the encoded structures of the 4-arc graph [12].

5. The spatial structure of the set of meridian cycles of the duplet genetic code can be represented by icosahedron (for details see below). The icosahedron and dodecahedron being related structures, one can apply the principles of anti-symmetry revealed in icosahedron to develop a system arranging the amino acids on the dodecahedron. Thus, taking into account all the underlined facts, we approached the problem of elaborating the model of spatial organization of the canonical set of amino acids.

2.2. *Meridian cycles of the hypercube B^4 and duplet genetic code*

It is known [17] that owing to the degenerate character of triplets the genetic code can be considered as a duplet one. Its spatial structure is described by Boolean hypercube B^4 [1, 3, 4], as represented in Figure 1a, based on the version of hypercube given in [18]. It should be noted that its dimension coincides with that of the physical space-time continuum. Symbols of duplets of bases placed in the vertices (circles) of the hypercube are connected by unit transitions and Boolean variables are assigned to the vertices according to the correspondence rules [12]: C = 00, U = 01, G = 10, A = 11.

Each vertex is also marked with symbols of amino acids encoded by the duplet corresponding to that particular vertex. First, we shall carry out the analysis and systematization of M-cycles on the hypercube B^4 and then extend the developed principles to the duplet code. Five levels can be distinguished in the structure of the hypercube B^4 (Figure 1,a). The first (top) and the fifth (bottom) levels consist of one vertex each. The second and the forth levels contain 4 vertices each and there are 6 vertices on the third level. The value of the Boolean variable corresponding to the first (top) vertex is a number containing four “0”, while that corresponding to the bottom vertex contains four “1”. Each vertex of the hypercube is connected by one bite transition with four neighboring vertices. The values of Boolean variables corresponding to symmetrically (left/right) situated vertices are symmetrical, e.g. 1000 and 0001, 0100 and 0010, etc.

There are eight duplets in the vertices (bold) situated on the upper three levels, which code for one amino acid each. Their corresponding Boolean variables can be converted into those of the lower level vertices by means of a

transformation 0 ↔ 1. A path leading from the vertex “0000” to the vertex “1111” and back makes M-cycle [16].

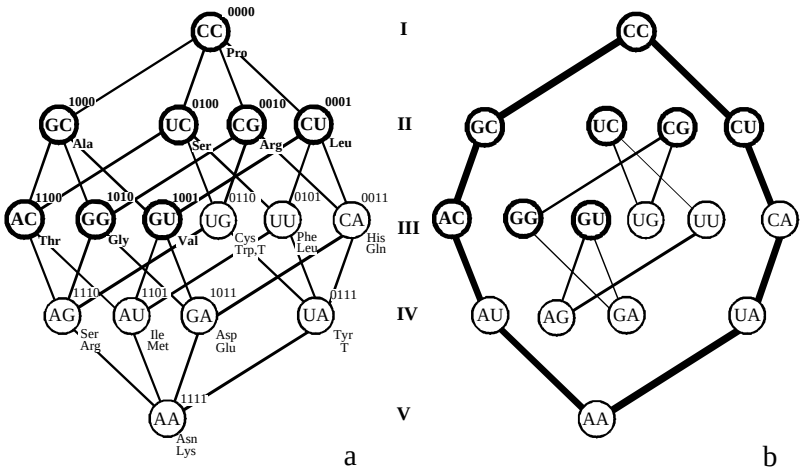
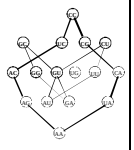
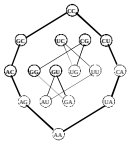
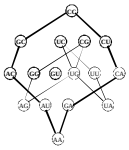
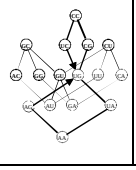
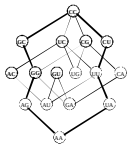
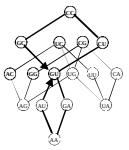
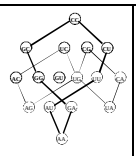
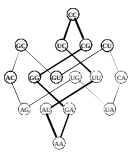
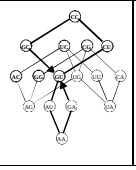
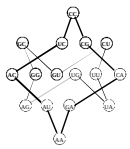
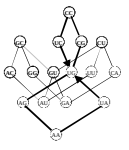


Figure 1. Structure of the duplet genetic code (a) and meridian cycle highlighted on this structure (b)

There are twelve M-cycles on the hypercube. Variables corresponding to the vertices, which form all these cycles, are shown in Table 2. As one can see in the table, left and right branches of all the cycles pass through the vertices with symmetrical values of variables, e.g. in the M-cycle 1 $_1$ these are 0100 – 0010, 1100 – 0011, 1110 – 0111. In the central column there are four cycles formed by the vertices so that the values corresponding to the upper part of one branch are anti-symmetric to those corresponding to the lower part of the other branch, e.g. in the cycle “1”: 1000 – 0111, 1100 – 0011, 0001 – 1110. In the left and right columns there are structures connected by a special relationship, which we called “mirror anti-symmetry”. Thus, in the cycle “1 $_1$” the upper pair of variables is represented by numbers 0100 – 0010, whereas variables of the lower pair in the cycle “ ^{1}1” assume values 1101 – 1011. One can easily observe this type of anti-symmetry by comparing the left number and the right number from the upper pair, respectively, with the right and left numbers from the lower pair, i.e. 0100 ↔ 1011 and 0010 ↔ 1101.

In addition, we have introduced another type of symmetry relationship, which we call rotational anti-symmetry. By definition, antipodes are such pairs

Table 2. Numeric representation of meridian cycles of Boolean hypercube B4 and topologies of that cycles in duplet genetic code.

Types of anti-symmetry		
Mirror (–)	Inner anti-symmetry	Mirror (+)
0000 **0100 0010** **1100 0011** **1110 0111** **1111** **1_1** 	**0000** **1000 0001** **1100 0011** **1110 0111** **1111** **1** 	**0000** **1000 0001** **1100 0011** **1101 1011** **1111** **^{1}1** 
0000 **0100 0010** **0110** **0111 1110** **1111** **2_1** 	**0000** **1000 0001** **1010 0101** **1110 0111** **1111** **2** 	**0000** **1000 0001** **1001** **1011 1101** **1111** **^{1}2** 
Rotational anti-symmetry		
0000 **1000 0001** **1010 0101** **1011 1101** **1111** **-1_1** 	**0000** **0100 0010** **0101 1010** **1101 1011** **1111** **-1** 	**0000** **0100 0010** **0101 1010** **0111 1110** **1111** **$-^1$1** 
0000 **1000 0001** **1001** **1101 1011** **1111** **-2_1** 	**0000** **0100 0010** **1100 0011** **1101 1011** **1111** **-2** 	**0000** **0100 0010** **0110** **1110 0111** **1111** **$-^1$2** 

of M-cycles, which have no variables in common, except for the initial and terminal, e.g. cycles "1_1" and "-1_1":

0000		**0000**		
0100 0010		**1000 0001**		
1100 0011		**1010 0101**		(1)
1110 0111	1_1	**1011 1101**	-1_1	
1111		**1111**		

Besides pairs given in (1), the following pairs "1" and " –1", "2" and " –2", "2_1" and "-2_1", "1^1" and "-1^1", "2^1" and "-2^1" also generate antipodes.

Let us consider M-cycles in the structure of the duplet genetic code. The types of cycles where pairs of numbers are encoded by letter symbols according to (1) are shown also in Table 2. They strictly correspond to the numeric

representation of cycles. Symmetric values are coded for by letters C and C, A and A, U and G. Letters C and A, G and U code for anti-symmetric values, which obey the principle stated by Rumer [17].

We can identify three groups of cycles: a) cycles with inner symmetry “1”, “2”, “ –1”, “ –2”; b) totally equivalent cycles “2 $_1$”, “ ^{1}2”, “ –2$_1$” and “ –^{1}2”; c) cycles with low symmetry “1 $_1$”, “^{1}1”, “–1$_1$”, “–^{1}1”. In Table 2 they are arranged according to their symmetry. Thus, cycles with inner symmetry are placed in the central column. Pairs of equivalent cycles and of cycles with low symmetry are situated symmetrically with respect to the central cycles (mirror anti-symmetry). Rotational antipode cycles from the second group occupy the second and the fourth line in the table, whereas the cycles from the third group occupy the first and the third line. Pairs of cycles “2 $_1$” and “–2$_1$”;“2^1” and “–2^1”, though looking similar, describe different paths.

2.3. *Icosahedron is a spatial structure of meridian cycles*

Let us place all M-cycles in 12 vertexes of the icosahedron (Figure 2), preserving the anti-symmetry relationship, fixed in Table 2. Plane I passes through the M-cycles possessing inner anti-symmetry, which are situated in the center of Table 2, and separates M-cycles, interrelated by operation of anti-symmetry suggesting changes in duplets according to the: C ↔ A, G ↔ U. This plane divides the cycles into two groups so that pairs cycles belonging to these groups are connected by mirror anti-symmetric transformation (0↔1).

As seen from Figure 2, plane II separates two groups of M-cycles, which do not have duplets in common other than CC and AA, the so called rotational antipode cycles. On rotating this plane about the axis perpendicular to plane I (C_2), vertices with antipode cycles coincide. Since icosahedron is related to dodecahedron, we applied the anti-symmetry principles established on the icosahedron to analyze the arrangement of amino acids on the dodecahedron.

3. The model

As has been established above (Section 2.1), icosahedron and dodecahedron are related structures. A question arises about the structural properties of the elements, which should be taken into account in order to arrange them on the dodecahedron preserving the symmetry elements identified in the icosahedron.

Let us denote elements lying in plane I of the dodecahedron (Figure 3) as A and B (compare with M-cycles 1 and 2 in Figure 2). Then rotational antipodes (–A), (–B) must lie in the same plane 1. Their position in the plane coincides

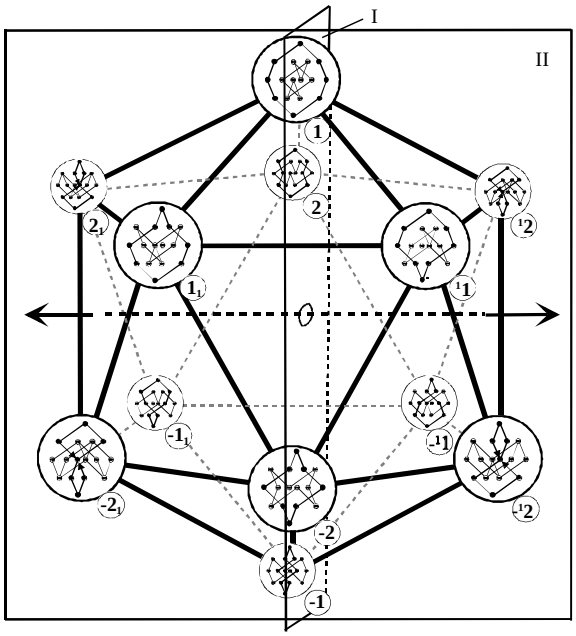


Figure 2. Localization of meridian cycles of the duplet genetic code on the icosahedron.
I – plane of mirror anti-symmetry; II – plane of rotational anti-symmetry and axis C_2.

with that of A, B upon rotation about axis C_2 (Figure 3). Beside each of these elements, two pairs of derivative mirror or quasi-mirror symmetric structures can be placed, e.g. for A these are structures A_1, ^1A and A_2, ^2A. These two pairs are complemented with two pairs of rotational antipode structures ($-A_1$), ($-^1A$) and ($-A_2$), ($-^2A$), situated beside the structure ($-A$). Similarly, a family of structures can be derived from the structures B and ($-B$). The two family of A and ($-A$) derivatives, regarded as vertexes, form, respectively, a top (dashed) and bottom (gray) pentagon facets. Other pentagon facets are formed by representatives of A and B structural lines. The total number of feasible structures is 20, which coincides with the number of amino acids composing the canonical set. Hence, if the developed scheme is suitable for the side chains of amino acids, then the canonical set of 20 amino acids should contain four simple amino acids (two antipode pairs) and eight pairs of derivatives, connected by quasi-mirror symmetry. On the other hand, the latter 16 amino acids form eight pairs of rotational antipodes.

We have applied the principles of anti-symmetry of the structures connected with M-cycles to systematization of the side chains of the amino acids from the

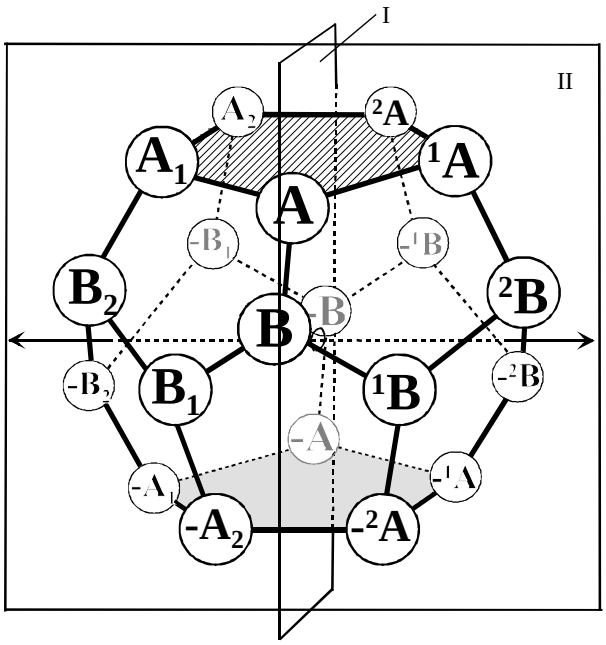


Figure 3. Representation on the dodecahedron of 20 elements reproducing symmetry properties of the icosahedron M-cycles.
I – plane of quasi-mirror symmetry; II – plane of rotational antipodes and axis C_2.

canonical set. The obtained results are shown in Figure 4.The side chains of all 20 amino acids can be divided into four groups: group of Pro, groups of Val, group of Gly, group of Ala. The side chains are united in groups on the basis of corresponding amino acids (placed in the center in the plane I), which are considered as the most simple. On the right and on the left from the central amino acid side chain there are quasi-mirror symmetric amino acids (His – Trp, Lys – Arg, Asn - Gln, etc). They are arranged symmetrically about the plane I.

In Pro- and Val-groups, the side chains of amino acids of smaller size (His, Phe, Asn, Lys) are situated left of the center, whereas amino acids of bigger size (Trp, Tyr, Gln, Arg), though having similar properties, lie on the right from the central amino acid. In Gly- and Ala-groups, which are rotational antipodes of amino acids from the first two groups, the order is different. Amino acids of bigger size (Thr, Met, Ile, Glu) are on the left, and those of smaller size (Ser, Cys, Leu, Asp) are on the right from the center. Hence, rotational antipode pairs

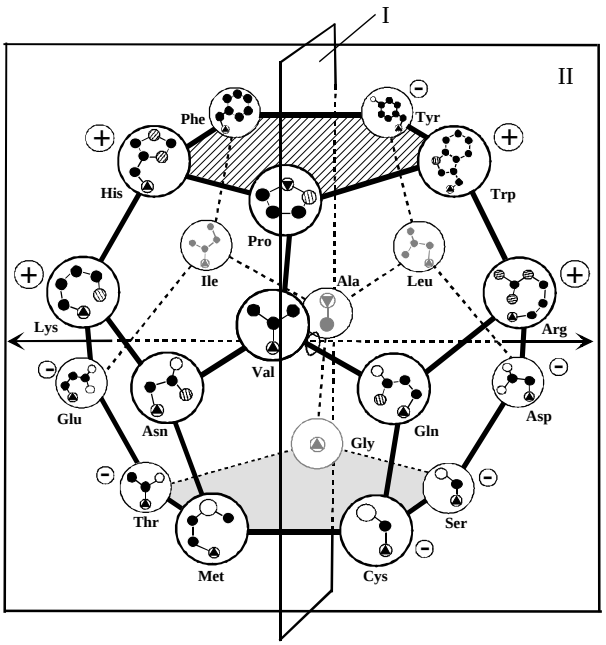


Figure 4. Dodecahedron like spatial structure of the canonical set of amino acids on the basis of symmetry relations.
I – plane of quasi-mirror symmetry; II – plane of rotational antipodes and axis C_2.

are formed so that the total molecular mass of the involved amino acids is minimized: His–Thr; Trp–Ser; Phe–Met and Tyr–Cys.

Pro- and Gly-groups form two pentagon cycles on the top (dashed) and on the bottom (gray), respectively. The groups of Val (five bigger circles in front) and Ala (five smaller circles, background) form a ring in the central part of the dodecahedron. Rotation of the plane II over 180° (C_2 axis) brings amino acids situated in front of the plane II (bigger circles) in coincidence with their antipodes situated behind the plane (smaller circles): His ↔ Thr, Phe ↔ Met, Trp ↔ Ser, Tyr ↔ Cys, Lys ↔ Glu, Arg ↔ Asp, Gln ↔ Leu, Asn ↔ Ile.

It should be noted that the side chains of rotational antipode amino acids possess opposite physicochemical properties, e.g. in the pairs Lys – Glu, Arg – Asp, His – Thr, the first amino acids exhibit proton donor, while the second ones - proton acceptor capacity. Neutral donor-acceptors Asn, Gln are antipodes of non-polar amino acids Ile, Leu. The side chain of Tyr, exhibiting donor-acceptor properties is an antipode of Cys. Neutral cyclic Phe is an antipode of neutral non- cyclic Met. It is interesting to notice that the side chains

of Ile and Leu are identical and the two amino acids differ only by the position of the α-carbon atom: in Ile it is on the bottom and in Leu it is on the opposite side.

4. Conclusion

In this paper a model of spatial structure for the arrangement of the canonical set of amino acids on the dodecahedron is proposed. A question arises about the adequacy of the proposed structure on to the intrinsic characteristics of amino acids. Arrangement of the amino acids belonging to particular group is structurally determined. Thus, in the pentagon cycle of Pro-group, equidistant and most close to Pro are Trp and His, both containing a five-member cycles and differing from Pro by an additional CH_2-group, which separates the five-member cycle from the main chain. Phe and Tyr are also cyclic amino acids, both containing a six-member cycle. They are separated from Pro by a two-step distance. The same is true for the pentagon cycle of Gly-group. Ser and Thr are one-step distant from Gly (CH_2OH-group is added), whereas Cys and Met are separated from Gly by two steps (O is replaced by S).

Similarly, one can analyze arrangement of amino acids in the central zone of the dodecahedron. These are amino acids related with Ala and Val via side chains of more complex structure. Obviously, the side chains of Lys and Arg; Asp and Glu; Asn and Gln as well as Ile and Leu belong to different groups of amino acids symmetrically situated about plane I. On the other hand, pairs Lys – Glu and Arg – Asp are clearly assigned to different antipode groups. We should also emphasize that changes undergone by the properties of amino acids within the pentagon cycles on the dodecahedron are gradual.

The proposed model has certain advantages in comparison with other approaches to the amino acids' side chains classification [5-12]. It is compatible with the fact, that the number of amino acids is 20. From the view-point of the model, it is specified by the number of vertices of the dodecahedron, 20, which serves topological basis for systematization of amino acids. It follows from the model that the conventional set of amino acids originates from an initial basic set of five amino acids multiplied by means of quasi-mirror reflection and rotation operations. The first operation generates pairs of amino acids of different structure but with similar properties, whereas the second operation produces pairs of antipode amino acids.

One of the feasible practical aspects of the investigated problem can be related to elaboration of molecular electronics devices. Specifically, it implies construction of an appropriate system of side chains as physical operators or

electronic modules [7, 19] with the main chain of different nature. The discovered principles of symmetry, which lie in the basis of the proposed classification system for amino acids, may contribute a new approach to the above technological problem. Feasibility of the molecular electronics model based on the polynucleotide chain as the main chain and modified nitrous bases as physical operators was considered in [19]. The dodecahedron-based model of 20 elements arrangement (Figure 3) presented in this paper may also serve prototype for constructing the system of modified bases and some other systems.

It may be expedient to take into account the developed principles in designing analogs of known compounds, e.g. tetrapeptides exhibiting biological activity. The model can be applied to the analysis of protein conformations emerging in the course of protein folding. The main advantage and distinction of the proposed model consists in its relative independence from other systems, e.g. such as the genetic code. In fact, structural performance of proteins is guided by its specific principles, determined by structural interrelations of amino acids. Future studies should demonstrate how productive the developed model for classification of amino acids is.

References

1. M.O. Bertman and J.R. Jungck, *J.Heredity* **70**, 379 (1979).
2. H.H. Klump, *Arch. Biochem. Biophys.* **301**, 207 (1993).
3. M.A. Jimenez-Montaño, C.R. de la Mora-Basañez and Th. Pöschel, *BioSystems* **39**, 117 (1996).
4. V.A. Karasev and S.G. Sorokin, *Russian J. Genetics* **33**, 622 (1997).
5. G.E. Schulz and R.H. Schirmer, *Principles of Protein Structure*, (Springer 1979).
6. P.N. Campbell and A.D. Smith, *Biochemistry Illustrated*, (Curchill Livingstone 1994).
7. V.A. Karasev, V.V. Luchinin and V.E. Stefanov, *Adv. Mater. Opt. Electron.* **4**, 203 (1994).
8. L.B. Mekler and R.G. Idlis, *Priroda*, **No.5**, 29 (1993).
9. J.E. Zull and S.K. Smith, *Trends Biochem. Sci.* **15**, 257 (1990).
10. G.M. Polya, *Nature* **255**, 652 (1975).
11. J.E. Blalock and E.M. Smith, *Biochem. Biophys. Res. Commun.* **121**, 203 (1984).
12. V.A. Karasev and V.E. Stefanov, *J. Theor.Biol.*, **209**, 303 (2001).
13. V.A. Karasev, *Deposited by VINITI*, **№470-B2004** (23.03.2004).
14. I. Hargittai and M. Hargittai, *Symmetry through the Eyes of a Chemist*, (VCH Verlaggesellschaft 1986).
15. V.A. Kolb, E.V. Makeyev and A.S. Spirin, *J. Biol Chem.* **275**, 16597 (2000).
16. V.A. Karasev, *Deposited by VINITI*, **№237-B2004** (10.02.2004).
17. Yu.B. Rumer, *Dokl. Akad. Nauk SSSR*, **183**, 225 (1968).
18. S.W. Yablonsky, *Basics of Discrete Mathematics*, (Nauka 1986).
19. V.A. Karasev, V.V. Luchinin, and V.E. Stefanov, *Adv. Funct. Mater.* **12**, 461 (2002).