

The naturally designed spherical symmetry in the genetic code

Chi Ming Yang

Physical Organic Chemistry and Chemical Biology, Nankai University, Tian Jin 300071, China

E-mail: yangchm@nankai.edu.cn

ABSTRACT

In the present work, 16 genetic code doublets and their cognate amino acids in the genetic code are fitted into a polyhedron model. Based on the structural regularity in nucleobases, and by using a series of common-sense topological approaches to rearranging the Hamiltonian-type graph of the codon map, it is identified that the degeneracy of codons and the internal relation of the 20 amino acids within the genetic code are in agreement with the spherical and polyhedral symmetry of a quasi-28-gon, *i.e.*, icosikaioctagon. Hence, a quasi-central, quasi-polyhedral and rotational symmetry within the genetic code is described. Accordingly, the rotational symmetry of the numerical distribution of side-chain carbon atoms of the 20 amino acids and the side-chain skeleton atoms (carbon, nitrogen, oxygen and sulfur) of the 20 amino acids are presented in the framework of this quasi-28-gon model. Two evolutionary axes within the 20 standard amino acids are suggested.

I. INTRODUCTION

The integrity, complexity and development mechanisms of the biological system form a puzzling subject. Wolpert and Lewis (1975) stated “The question of how these component parts (including the genetic networks) are organized into a complete control system for development is posed as a problem for future study”. Particularly, the genetic code has been a conceptual challenge to all scientists since it was cracked down four decades ago. Towards understanding the origin and evolution of the genetic code, a few early landmarks include Crick’s discussion about two fundamentally different theories of genetic code evolution, *i.e.*, “frozen accident theory” and the “stereochemical principle” or the “natural selection and co-evolution theory” (Crick, 1968; Wong, 1975), and an illuminative hypercycle principle from Eigen (Eigen, 1971; Eigen and Schuster, 1977; 1978; 1979; Eigen et al., 1981; Blomberg, 1997; Ycas, 1999).

Since Woese introduced a legend observation that the four RNA nucleobases, uracil (U), cytosine (C), adenine (A) and guanine (G) can be doubly ordered by ranking them according to their relative potentials as electron donors as well as their relative polarity/hydrophobicity (Woese et al., 1966; Woese, 1965; Woese, 1967), based on which the error minimization in the genetic code can be quantitatively measured (Haig and Hurst, 1999), it is now widely realized that genetic codons and amino acids can be shown to be related by chemical principles (Siemion and Stefanowicz, 1992; Rodin et al., 1993; Di Giulio et al., 1994; Cedergren and Miramontes, 1996; Di Giulio and Medugno, 1998; Knight and Landweber, 1998; Yarus, 1998; Knight et al., 1999; Yarus, 2000; Di Giulio, 2001; Di Giulio and Medugno, 2001). Presently, the internal regularity displayed by the genetic code has been recognized to include physicochemical property correlation, biosynthetic relation of amino acids and the non-random pattern of the genetic code.

Within the past few decades, accompanied with the codon origin and evolution of the genetic code which have fascinated scientists across numerous disciplinary fields, the inherent symmetry characteristics of the genetic code remains another unresolved but intriguing subject. Despite excellent efforts and innovative interdisciplinary approaches examining the unclear symmetry feature of the genetic code, however, a precise description of the symmetry characteristics inherent in the

standard genetic code including, in particular, why the total number of standard amino acids is 20, still awaits a clear elucidation.

The 20 standard amino acids selected in the genetic code constitute a paradigm of complexity in Nature's integrity (Dufton, 1997; Davydov, 1998). Although biological importance of the code suggest a system complexity within the code, a multi-disciplinary in-depth examination of the code described here show its basic symmetric feature has great simplicity. In a previous paper (Yang, 2003) we presented a line of reasoning favoring the order of bases as UCGA succession in listing the genetic code. We used sp^2 N-atom number to rank nucleobases and the resulted genetic code shows a nice correlation between amino acid property and sp^2 N-numbers. The purpose of this work is to describe the newly identified symmetric three-dimensional relation of the 20 amino acids within the genetic code. This finding may have novel features, which are of considerable biological interest.

II. METHOD AND RESULTS

1) An empirical stereo-electronic property of nucleobases for a rearranged genetic code

Chemical structures of the four nucleobases are mainly a six-membered ring, with or without infusion to a five-membered ring, in which all the heteroatoms (O, N) are in a conjugated position via their covalent bond linkages. Thus RNA nucleobases themselves display certain types of molecular structural regularity. I recently took one measure of this molecular regularity, covalent bonding hybrid of nitrogen atoms, as a determinative measure of chemical structural property to further discriminate among the mRNA nucleobase A, U, G and C. The sp^2 N-atom number in these nucleobases are 3 for A, 2 for G, 1 for C and 0 for U, respectively (Figure 1). If using this set of sp^2 N-atom numbers in the nucleobases as a determinative measure, a rearranged code is obtained in Table 1 and Figure 2, the latter is listed in a three-dimensional space.

Early studies demonstrated that a rearranged genetic code could sometimes be unexpectedly informative (Grantham, 1980; Jiménez-Montaña et al., 1996; Jiménez-Montaña, 1999; D'Onofrio et al., 1999; Szathmary; 1999; Lehmann, 2000). RNA nucleobases listed in the succession of UCGA is coincidentally consistent with nucleobase hydrophilicity values from paper chromatography, which was invariably $A < G < C < U$ (Weber and Lacey, 1978). Therefore, the genetic code listed in the order of AGCU, *i.e.*, increasing mononucleotide hydrophilicity in a correlation with increasing Woese's amino acid polar-requirement (Woese et al., 1966; Woese, 1965; Woese, 1967), was employed by Grantham (1980), Jimenez-Montano (Jiménez-Montaña et al., 1996; Jiménez-Montaña, 1999), D'Onofrio and co-workers (D'Onofrio et al., 1999) Szathmary (1999) and Lehmann (2000) and in their studies.

Number of sp^2 nitrogen atoms in RNA nucleic bases: A, G, C and U

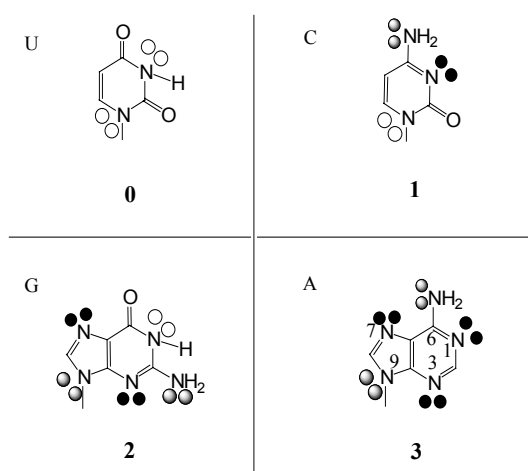


Figure 1. Nucleobases display unique chemical structure regularity. The four RNA nucleobases carry different number of sp^2 N atoms: 3 for A, 2 for G, 1 for C, to 0 for U, respectively. The nitrogen atom that carries one lone-electron-pair (LEP) (“••”) is a sp^2 hybrid nitrogen-atom.

Table 1. A slightly rearranged codon map, where the four nucleobases forming tri-nucleotide codons are placed in the order of U, C, G and A.

1 st letter \ 2 nd letter	U (aa)	C (aa)	G (aa)	A (aa)
U	UUU (000) <i>F</i> UUC (001) <i>F</i> UUG (002) <i>L</i> UUA (003) <i>L</i>	UCU (010) <i>S</i> UCC (011) <i>S</i> UCG (012) <i>S</i> UCA (013) <i>S</i>	UGU (020) <i>C</i> UGC (021) <i>C</i> UGG (022) <i>W</i> UGA (023) <i>Stop</i>	UAU (030) <i>Y</i> UAC (031) <i>Y</i> UAG (032) <i>Stop</i> UAA (033) <i>Stop</i>
C	CUU (100) <i>L</i> CUC (101) <i>L</i> CUG (102) <i>L</i> CUA (103) <i>L</i>	CCU (110) <i>P</i> CCC (111) <i>P</i> CCG (112) <i>P</i> CCA (113) <i>P</i>	CGU (120) <i>R</i> CGC (121) <i>R</i> CGG (122) <i>R</i> CGA (123) <i>R</i>	CAU (130) <i>H</i> CAC (131) <i>H</i> CAG (132) <i>Q</i> CAA (133) <i>Q</i>
G	GUU (200) <i>V</i> GUC (201) <i>V</i> GUG (202) <i>V</i> GUA (203) <i>V</i>	GCU (210) <i>A</i> GCC (211) <i>A</i> GCG (212) <i>A</i> GCA (213) <i>A</i>	GGU (220) <i>G</i> GGC (221) <i>G</i> GGG (222) <i>G</i> GGA (223) <i>G</i>	GAU (230) <i>D</i> GAC (231) <i>D</i> GAG (232) <i>E</i> GAA (233) <i>E</i>
A	AUU (300) <i>I</i> AUC (301) <i>I</i> AUG (302) <i>M</i> AUA (303) <i>I</i>	ACU (310) <i>T</i> ACC (311) <i>T</i> ACG (312) <i>T</i> ACA (313) <i>T</i>	AGU (320) <i>S</i> AGC (321) <i>S</i> AGG (322) <i>R</i> AGA (323) <i>R</i>	AAU (330) <i>N</i> AAC (331) <i>N</i> AAG (332) <i>K</i> AAA (333) <i>K</i>

Amino acids (aa's) and their codons in blue and green color form a crossed intersection. Abbreviations of the 20 amino acids are represented by A(Ala), P(Pro), V(Val), G(Gly), T(Thr), S(Ser), L(Leu), R(Arg), D(Asp), E(Glu), M(Met), I(Ile), F(Phe), C(Cys), W(Trp), H(His), Q(Gln), N(Asn), K(Lys) and Y(Tyr).

2) A three-dimensional display of the 16 genetic code doublets in the rearranged genetic code

We now investigate the symmetry feature starting from a three-dimensional display of the rearranged genetic code. Given the 64 genetic codes comprising of 16 genetic code doublets, the strong group feature of the genetic code was initially recognized by Bertman and Jungck (1979), these genetic code doublets can be divided into two octets of completely degenerate and ambiguous coding dinucleotides. Findley and co-workers (1982) used a similar approach as that of Gatlin (1972), reasoned that the genetic code is a relation rather than a mapping by using empirical arguments processed within a group-theoretic framework (Findley et al., 1982; Gatlin, 1972). Findley discussed in particular the inherent symmetry as the even-order degeneracy constraint together with the odd-order degenerate codons in the genetic code.

Based on all of these work, the genetic code is now three-dimensionally presented in Figure 2, which more clearly display the internal relation between each two groups of the 16 genetic code doublets, with every line connecting two genetic code doublets, which vary by one base-letter from one to another.

To unravel the concealed symmetry feature within the genetic code, a series of topological approaches to rearranging the codon map is carried out. First, using a simple and common topological sense, the 3-D map in Figure 2a can be depicted by the following Hamiltonian graph in Figure 2b. In this graph, the internal relation between the 16 genetic code doublets are illustrated, each one of 16 genetic code doublets being connected with four other genetic code doublets, this connectedness corresponding to one base letter change between two neighboring genetic code doublets in the rearranged genetic code.

3) Rotational symmetry revealed from further topological transformation of the genetic code

To elucidate the hidden symmetry inherent in the genetic code, I further topologically rearrange the genetic code in a step depicted in Figure 3, to reach a closed spherical graph. Based on the discussion in the proceeding section and by placing the possible codon core on the bottom of the sphere, another way deciphering the 3-dimensional codon map is hence obtained in Figure 3b. A spherical shape in this display not only exhibits rotational symmetry and spherical feature of the genetic code system, but also graphically, in a visual sense, explains why the genetic code system has the capability of self-maintaining its integrity.

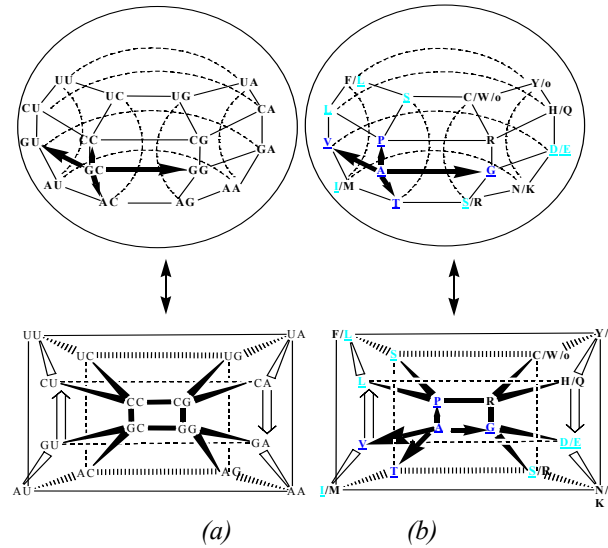


Figure 2. Reaching a Hamiltonian-type graphic display of the 16 genetic code doublets and their amino acids. Each vertex is connected with four other vertices and each vertex represents a genetic code doublet corresponding to its cognate amino acid(s): a) The 16 genetic code doublets of the rearranged genetic code in a three-dimensional space, whereby the four nucleobases are placed in the succession of UCGA. b) A three-dimensional display of the 20 amino acids corresponding to the 16 genetic code doublets. Four black arrows illustrate the core presumably formed at a certain evolution stage by five groups of genetic code doublets (amino acids are “A, P, V, G and T”) (Yang, 2003). Amino acids in both blue and light-blue color are produced from Miller's prebiotic simulation experiment (Weber and Miller, 1981). “o” denotes “stop” codons. Abbreviations: 20 amino acids are represented by A(Ala), P(Pro), V(Val), G(Gly), T(Thr), S(Ser), L(Leu), R(Arg), D(Asp), E(Glu), M(Met), I(Ile), F(Phe), C(Cys), W(Trp), H(His), Q(Gln), N(Asn), K(Lys) and Y(Tyr); aa(amino acid).

(*Note 1:* We have previously re-classified the 20 amino acids into 5 groups of structures, according to their stereochemistry, which overlap precisely the 5 genetic code doublets in the crossed-intersection in a), it was suggested a codon core may have been formed at the primordial stage (16).)

(*Note 2:* A Hamilton cycle is a cycle that includes each vertex exactly once (in other words, it is a spanning cycle). Since, if a map has a Hamilton cycle, then it can be four-colored (or 4-colorable) (Figure 3). A 4-colorable genetic code map is consistent with a requirement of four RNA bases in the whole genetic codes, in which the triplet (nucleotide) feature of codons reflects 3 dimension-character of each codon, as is displayed by each of the three base letters in every tri-nucleotide codon.)]

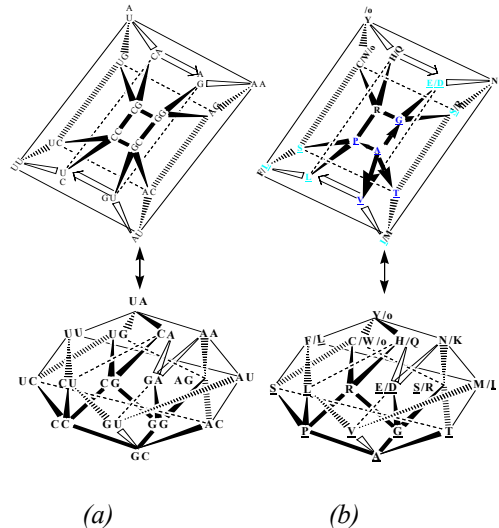


Figure 3. From Hamiltonian-type graph in Figure 2 to a closed spherical display which reveals the self-maintenance ability and the integrity of the genetic code. The 64 codons can be divided into 16 groups, half of that encode a single amino acid, having codon redundancy of 4. All the 3 sextets in degeneracy (each of the 3 amino acids encoded by 6 codons) are in the middle regions. Among the 5 amino acids forming a pyramid “A (P,V,G,T)”, each amino acid is uniformly encoded by 4 codons from one genetic code doublet. The pyramid A (P,V,G,T) region is consistent with the early vision of primordial genetic code. Amino acids underlined are produced from Miller's simulation experiment (Weber and Miller, 1981). “o” means “stop” codon. The map shows that the more complex functionality an amino acid side-chain carries, the further away the amino acid is from Ala codons. If Ala is designated as the “Southern Pole”, Tyr /o is the “Northern Pole”. In this map, all the aromatic amino acids (Y, F, W, H) and stop codon “o” are in the same region, reminiscent of the current codon range expanding effort that is now being widely practiced in the “Tyr/o” codon region.

4) A comparison of the two three-dimensional map based on two genetic code table

Thus, when the arranged genetic code topologically presented in arrangement (II) on the basis of U, C, G and A succession, is compared with the commonly used genetic code presented in arrangement (I) whereby nucleobases are listed in a succession of U, C, A and G, it is evident that the new arrangement (II) from the rearranged genetic code (Table 1) can better display the high rotational symmetry in coding degeneracy, see Figure 4.

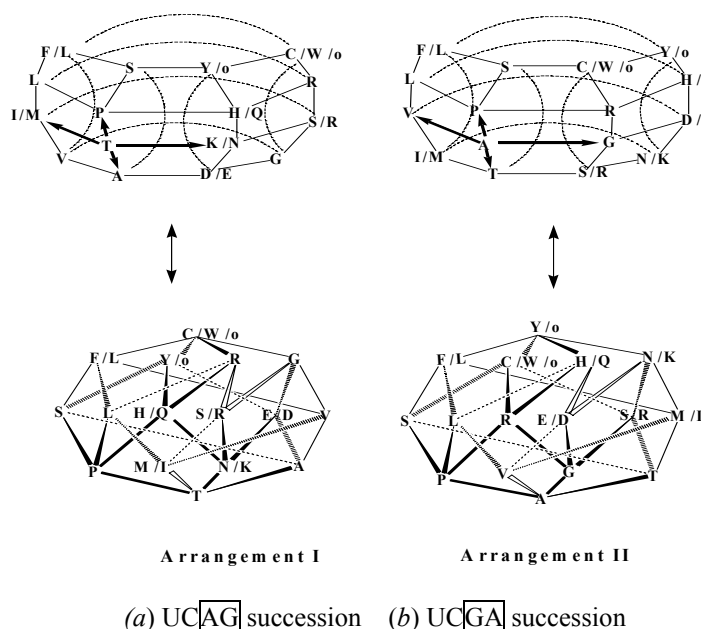


Figure 4. Two types of “world-map” arrangement of the 20 amino acids corresponding to two types of closed spherical representation of the 16 genetic code doublets: I) Nucleobases in the corresponding genetic code are placed in the succession UCAG; II) Nucleobases in the corresponding genetic code are placed in the succession UCGA

5) A polyhedral model (quasi-28-gon) helps reveal hidden symmetry of the genetic code

(1) Summarization of supersymmetry of the rearranged genetic code in the UCGA succession using a polyhedral model

Now, I select the code in arrangement II for a further discussion. Consider both the closed spherical feature with the newly identified rotational symmetrical feature of the code, the genetic code in arrangement II can be conveniently summarized by a polyhedron model in order to more definitely describe the internal symmetric relation of 16 genetic code doublets and their 20 cognate amino acids (Figure 5). In this model, symmetry in coding degeneracy can be described using a quasi-28-gon. In Table 2, one could amplify and generalize some of concepts.

The consequently elucidated amino-acid assignment and distribution around a quasi-28-gon comply with the general even-order degeneracy constraint (the fundamental degeneracy is of order 2 in genetic code), which is the basic symmetry as defined for the doubly degenerate codons. In addition to order-4 and order-6 degenerate codons in the genetic code, there are two sets of triply-degenerate codons, one of which maps onto Ile while the other maps onto Stop, and two nondegenerate codons, one of which maps onto Met while the other maps onto Trp. A quasi-28-gon helps clearly indicate that slight deviations from strict symmetry have occurred at the Y/o, C/W/o and M/I genetic code doublet positions. Despite these odd-order degenerate codons, nevertheless, the total number of amino acids at these positions remains C-symmetrical (Table 2). Notably, the “o” (stop) codons are not totally non-sense, but allow a counterbalance for numerical distribution of both amino-acid and side-chain C-atom along a presumed evolutionary axis (Table 2, 3 and Figure 7).

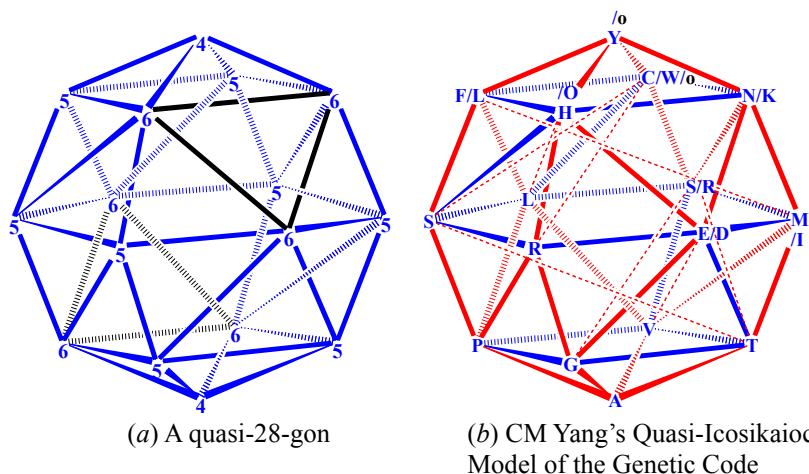


Figure 5. Pattern of symmetry in the genetic code displayed in a polyhedron (quasi-icosikaioctagon or quasi-28-gon). (a) The basic symmetry elements in a quasi-28-gon include 6 C_2 axes; 1 C_2 symmetric plan and 1 I (one symmetric center). The number at each vertex indicates the number of edges. (b) A solid geometrical illustration of the 3-dimensional codon map using a quasi-28-gon (Icosikaioctagon), displaying a highly symmetrical evolution of both the 20 amino acids and the 16 genetic code doublets. The letter(s) at each vertex denotes amino acid(s) coded by one of the 16 genetic code doublets. This map explains “why 20”. The distribution of the 20 amino acids is symmetrical along a presumed evolutionary axis from Ala codons to Tyr codons: the number of the genetic code doublets (number of amino acids) at every stage starting from Ala codons to Tyr codons are: 1(1); 4(1,1,1,1); 6(1,1,1,2,2,2); 4(2,2,2,2) and 1(1,0), respectively. In (b), block lines (both red and blue) are edges for polyhedron construction; red lines (both block and dotted) are for neighboring code-doublet connection.

Note 3: For any map on a given surface, if the number of vertices of such a map is v , the number of arcs (or edges) is e , and the number of regions (or faces) is f , then, the number obtained from the expression $v - e + f$ is denoted by x , and is termed the Euler characteristic of the surface. For a sphere, $x = 2$. Any polyhedron, which is both closed and convex, is termed a simple polyhedron. Such a polyhedron may be continuously deformed into a sphere, and hence it follows that $v - e + f = 2$. For Icosikaioctagon or 28-gon of the genetic code: $x = v - e + f = 16 - e + 28 = 2$; therefore, $e = 42$.

Table 2. Symmetry in the degeneracy (numerical distribution) of the 20 amino acids and 16 genetic code doublets revealed by a 3-D code map (arrangement II in Figure 4) or the polyhedron model (Fig. 5).

Trinucleotide codons and aa's	Total number of aa's	Trinucleotide codons and aa's	Total number of aa's	Trinucleotide codons and aa's	Total number of aa's
		UAN Tyr, "Stop"	1		
UUN Leu, Phe	4	↑↑ A presumed evolutionary axis of codons		CAN His, Gln	4
UGN Cys, Trp, "Stop"				AAN Asn, Lys	
UCN Ser	3			AGN Arg, Ser	6
CUN Leu				GAN Asp, Glu	
CGN Arg				AUN Ile, Met, "Start"	
CCN Pro	2			GGN Gly	2
GUN Val				CAN Thr	
		GCN Ala	1		

Note 4: N = U, C, G and A. Amino acids in black color are coded by codon from two different genetic code doublets, i.e., Sextets (Quartets + Duets): Arg; Leu and Ser. Degeneracy in the genetic code includes,

Even numbers:

3 Sextets (Quartets + Duets):

Arg; Leu; Ser;

5 Quartets (Quadruplets):

Thr; Pro; Ala; Gly; Val;

9 Duets:

Lys; Asn; Gln; His; Glu; Asp; Tyr; Cys; Phe;

Odd numbers:

2 Triplets:

Ile; terminators (STOP)

2 Singlets:

Met, Trp

(2) Cooperative symmetry of the numerical distribution of side-chain C-atoms of the 20 amino acids

Now we count the side-chain carbon atoms of amino acids at each genetic code doublets. The cooperative symmetry in their distribution is then summarized in Table 3, and Figure 7.

Table 3. Symmetry in the numerical distribution of side-chain C-atoms of the 20 amino acids revealed a 3-D code map (arrangement II in Figure 4) or the polyhedron model (Fig. 5)

Amino acids (C-atom number on side chains)	Total number of C-atoms on side chains	Amino acids (C-atom number on side chains)	Total number of C-atoms on side chains	Amino acids (C-atom number on side chains)	Total number of C-atoms on side chains
		UAN Tyr(7), o	7		
UUN Leu(4), Phe(7) AAN Asn(2), Lys(4)	17	↑↑ A presumed evolutionary axis of codons		CAN His(4), Gln(3) UGN Cys(1), Trp(9)	17
UCN Ser(1) GAN Asp(2), Glu(3) AUN Ile(4), Met(3), o	13			AGN Arg(4), Ser(1) CUN Leu(4) CGN Arg(4)	13
CCN Pro (3) ACN Thr(2)	5			GGN Gly(0) GUN Val(3)	3
				GCN Ala(1)	1

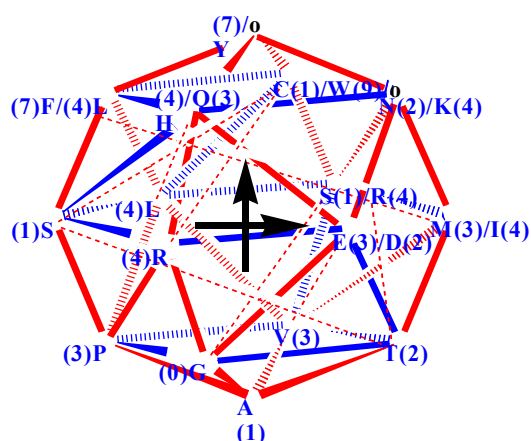


Figure 6. Number of carbon atoms on the side chains of amino acids. It is shown that any amino acids carrying a bigger number of side-chain C-atoms (>2) can be the summation of two amino acids carrying a smaller number of side-chain C-atoms, presumably consistent with a concerted stepwise coevolution of the canonical amino acids and their codons. Two possible evolutionary axes are indicated by arrows “↑” and “→”.

Further information can be obtained from the quasi-28 model in Figure7. We illustrate both the rotational symmetric feature of amino-acid coding degeneracy and symmetrical distribution of the sub-total number of C-atoms on side chains of amino-acid (s) at each vertex (genetic code doublet) based on a quasi-28-gon model. Moreover, results show that any amino acids carrying a bigger number of side-chain C-atoms (>2) can be the summation of two other amino acids carrying a smaller number of side-chain C-atoms, presumably consistent with a concerted stepwise evolution of the canonical amino acids and their codons---from simple to complex. Two evolutionary axes are thus proposed and indicated by arrows “↑” and “→”.

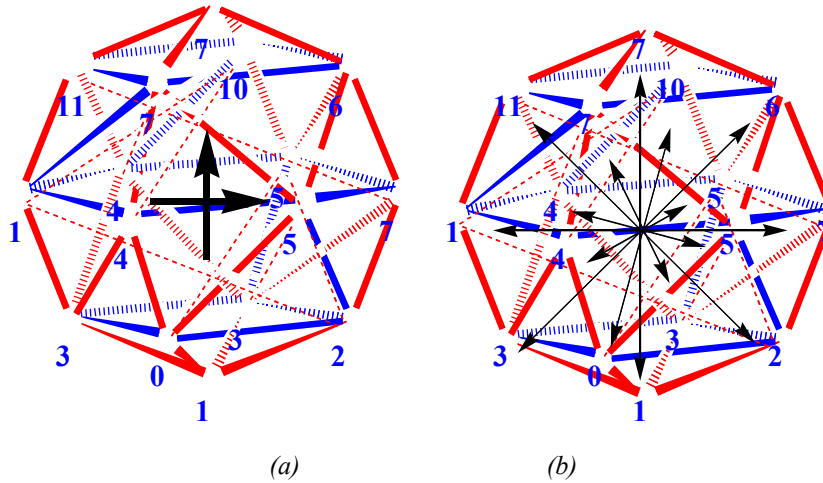


Figure 7. (a) Summarization of the sub-total number of C-atoms on side chains of amino acid(s) at each vertex (genetic code doublet) based on a quasi-28-gon model. Two proposed evolutionary axes are indicated by “↑” and “→”. (b) C₂-symmetry displayed in the sub-total number of side-chain C-atoms of amino acid(s) coded by two genetic code doublets at the opposite vertices, which are listed in groups of two perpendicular vertex-connection lines (+): 8, 8; 9, 9; 10, 10 and 9, 13.

(3) Cooperative symmetry in the numerical distribution of side-chain skeleton (C, N, O, S) atoms of the 20 amino acids

The rotational symmetry, along the presumed evolutionary axis from Ala codons to Tyr codons, is also displayed as a cooperative symmetry in the numerical distribution of sub-total number of side-chain C(carbon)/N(nitrogen)/O(oxygen)/S(sulfur) atoms of the 20 amino acids based on a polyhedron model, see Table 4.

Table 4. Symmetry in the numerical distribution of sub-total number of side-chain C/N/O/S atoms of the 20 amino acids based on a polyhedron model.

Amino acids (C/N/O/S-atom number on side chains)	Sub-total number of C/N/O/S-atoms on side chains	Amino acids (C/N/O/S-atom number on side chains)	Total number of C/N/O/S-atoms on side chains	Amino acids (C/N/O/S-atom number on side chains)	Sub-total number of C/N/O/S-atoms on side chains
		UAN Tyr(8), o	8		
UUN Leu(4), Phe(7)	11	↑↑ A presumed evolutionary axis of codons		CAN His(6), Gln(5)	11
AAN Asn(4), Lys(5)	9			UGN Cys(2), Trp(10), o	12
UCN Ser(2)	13			AGN Arg(7), Ser(2)	26
CUN Leu(4)				GAN Asp(4), Glu(5)	
CGN Arg(7)				AUN Ile(4), Met(4), st	
CCN Pro (3)	6			GGN Gly(0)	3
ACN Thr(3)				GUN Val(3)	
		GCN Ala(1)	1		

III. DISCUSSION

Early graphic approaches to understanding the genetic code include a genius display of the genetic code by employing a hypercube structural presentation by Jimenez-Montano et al. (1996), a codon graph by Bertman and Jungck (1979), a geometric model study by Halitsky (1994) and a topological model by Karasev and Sorokin (1997). However, the precise symmetry features within the relation of the 20 amino acids have remained elusive. Now, if one uses a polyhedron to illustrate the genetic code table 1, we can have a clearer view of the spherical relation within the 16 genetic code doublets and the 20 amino acids in the genetic code. Upon an integrated approach to looking at the geometry, symmetry, chemistry and topology of the genetic code, it was identified that both the distribution of the 20 amino acids and degeneracy of codons are in agreement with the cooperative

symmetry of a polyhedron 28-gon (or icosikaoctagon, Figure 5).

From this polyhedron model, we can find that the symmetric distribution of the 20 amino acids around the 28-gon excellently fits Rumer's regularity and explain the arithmetical regularity discovered by Shcherbak (1993). Both the amino acid in Group IV (degeneracy 4) and the rest of the amino acid within Quasi-group III-II-I (degeneracy 1; 2 and 3) are occupying half of the 28-gon (Figure 6).

Therefore, a general symmetric feature of the code described by our spherical and polyhedral approach together with two evolutionary axes is not only consistent with previous work, but also in an excellent agreement with the recent Trifonov's proposal that Ala could be the first codonic amino acid in the code that has been reached previously by quite other reasoning (Trifonov and Bettecken, 1997; Trifonov, 2000). The conclusion that a few amino acids and their codons are initially formed in the primordial stage was also envisaged by numerous previous works (Siemion and Stefanowicz, 1992; Rodin et al., 1993; Di Giulio et al., 1994; Cedergren and Miramontes, 1996; Di Giulio and Medugno, 1998; Knight and Landweber, 1998; Yarus, 1998; Knight et al., 1999; Yarus, 2000; Di Giulio, 2001; Di Giulio and Medugno, 2001).

The degeneracy in the standard genetic code has been shown to be ambiguous under certain condition, that is, more than one amino acid appears to correspond to the same codon, nevertheless, it has long been postulated that distribution of the total number of amino acids and "why 20" shall follow a symmetric pattern with the codon table (Weber and Miller, 1981). So far, symmetry and coding properties of genetic code has received tremendous attention (Portelli and Portelli, 1985; Shcherbak, 1988; Chipens et al., 1989; Shcherbak, 1989; Chipens, 1991a; 1991b; Hornos and Hornos, 1993; Arquest and Michel, 1996; Koch and Lehmann, 1997; Bashford et al., 1998; Frappat et al., 1999; Hornos et al., 1999; Bashford and Jarvis, 2000; Frappat et al., 2001; Lacan and Michel, 2001; Balakrishnan, 2002), and the pattern and symmetries inherent in the genetic code has been subjected to a wide variety of investigation approaches, including group theoretical analysis such as Lie algebra (Halitsky, 1994; Bashford et al., 1998; Hornos et al., 1999; Bashford and Jarvis, 2000; Balakrishnan, 2002;), hypercube structural presentation (Jimenez-Montano et al., 1996)), codon graph group (Bertman and Jungck, 1979) and geometric model study (Halitsky, 1994; Karasev and Sorokin, 1997). A codon ring as a theoretical model showed that the genetic code is almost an example of a Gray code (Swanson, 1984). The genetic code was more recently shown as a binary code, determined by Golden mean through the unity of the binary-code tree and the Farey tree (Rakocevic and Jokic, 1996).

Based on the quasi-28-gon model, it is identified herein not only the codon degeneracy is rotationally symmetric along a presumed evolutionary axis from Ala to Tyr codons, but also the side-chain C-atom numbers as well as the side-chain C/N/O/S atom-numbers are distributed in a rotationally symmetric way along this evolutionary axis (Table 4). This finding may be of interest to the current code expanding effort (Döring et al., 2001; Wang et al., 2001).

Although it has been known the genetic code is optimized and fixed (Freeland and Hurst, 1998; Freeland et al., 2000), its contents, in particular, the fundamental but unclear physicochemical property of 20 amino acids together with the distribution of amino acids in the genetic code has since inspired tremendous amount of excellent research and speculation including from Weber and Miller (1981), Swanson (1984), Luo (1989), Shcherbak (1993), Dufton (1997) and Davydov (1998). When Rakočević and Jokic (1996) suggested that the genetic code is determined by Golden mean, it was identified that atom number balance in amino acids are directed by Golden mean route, also directed by the double-triple system of amino acids, as well as by two classes of enzymes aminoacyl-t-RNA synthetases. Rosen (1999) described that the codon range numbers that follow for the 20 amino acids are shown to be given by linear Diophantine and explicit molecular content equations in the number of carbon, nitrogen, oxygen and sulfur atoms in each amino acid, suggesting that the universal genetic code that associates codons and amino acids is expressed in a precise way by purely physical molecular content relations.

To the best of my knowledge, it has for the first time been identified that the polyhedral symmetries as a quasi-28-gon are inherent in the genetic code. The spherical and polyhedral feature in the genetic code presented here is aimed to understand not just what now is, but the ways what now is might plausibly be expected to have arisen (Kauffman, 1993).

In summary, the rotational and spherical symmetry of codon-amino acid relation are displayed as

follows:

- a. According to our quasi-28-gon model, there is rotational symmetry and a deviated symmetric center within the code, but this symmetry is slightly destroyed by evolution in the code;
- b. The long-conjured co-operative symmetry character in the genetic code can be more precisely described by using a quasi-28-gon;
- c. The circular property of the code and hypercycle theory of the genetic code can be further explored by using the quasi-28-gon model;
- d. Amino acids distribution within the genetic code is symmetric along a possible evolutionary axis; This symmetry may have been established in the early evolutionary stage, in which A, P, V, G and T may have formed a possible core (Yang, 2003), consistent with a notion that Ala codons may be the first amino acid and codons;
- e. The 16 genetic code doublets from the 64 codons can be divided into 5 stages along a presumed evolutionary axis (Figure 6). Those 5 stages contain 1 group, 4 groups, 6 groups, 4 groups, 1 group of genetic code doublets, respectively. The numbers of genetic code doublets and the number of amino acids encoded at every stage are: 1(1); 4(1,1,1,1); 6(1,1,1,2,2,2); 4(2,2,2); 1(1,0). Exactly half of the 16 genetic code doublets are encoding 2 amino acids within each genetic code doublets except at Tyr position, and occupying half of the 3-D codon map. Alternatively, as revealed in the polyhedral model, the 20 standard amino acids and their codons would be stepwise evolutionarily divided into four groups, which may reflect their chronological order: A > P, V, G, T > S, L, R, D, E, M, I > F, C, W, H, Q, N, K > Y. Therefore, the polyhedron model, quasi-28-gon, supports our early suggestion that an early “frozen core” may have been formed in the primordial stage to attain 5 codon quartets for A and P, V, G and T, which may have inherent tendency in shaping the rest of genetic code (Yang, 2003).
- f. Results from the quasi-28-gon model are in agreement with a coevolution theory of the genetic code. Since code and amino acids are biochemically and/or stereochemically linked, as evident from Figure 4b and 5b, amino acid evolutionary trend along the presumed evolutionary axis in this model is consistent with chemical evolution, e.g., from D/E to Q/N; F to Y and S to C;
- g. Given the current interest in code range expansion (Döring et al., 2001; Wang et al., 2001), the quasi-28-gon model for the code reported herein is especially worthy of our attention. The polyhedral symmetry corresponding to quasi-28-gon helps understand the logic underlying why incorporating Tyr-derivatives (O-methyl-L-tyrosine, photocrosslinking amino acid) into amber codon (UAG) will not break the rotational symmetry along the presumed evolutionary axis, from Ala codons to Tyr codons (Figure 3b, 5b), within the genetic code. Similarly, when amber codons in the methyltransferase genes of certain archaea encode pyrrolysine as the 22nd amino acid, the rotational symmetry is little affected. However, when selenocysteine, identified as the 21st amino acid (in 1986), was directly encoded by UGA, which otherwise usually specifies translation termination (stop codon), actually breaks the rotational symmetry along the presumed evolutionary axis within the genetic code.

Since spherical symmetry is always associated with polyhedral symmetry, therefore, the polyhedron relation of the genetic code with features of rotational symmetry is not merely a natural accident. Icosahedral symmetry, which is a type of rotational symmetry, is regularly encountered, described as *Ih* symmetry. This is a common structure in virus coats, or capsids. A number of viruses such as poliovirus with its genome of RNA, have a protein capsid with a structural symmetry of icosahedron (Racaniello, 1996; Vargas et al., 1999; Lanzavecchia et al., 2002). The quasi-28-gon model may immediately suggest, in addition to the evolutionary logic currently tested by code-expanding effort, a possible mechanism in RNA translation. Therefore, it may not escape our notice that the quasi-28-gon symmetry in the genetic code with two “poles” as Ala codons and Tyr codons may have multiple biological implications, including the complex protein-synthesis mechanism which imposes another challenging problem for investigation (Liljenstrom and Blomberg, 1987; Krakauer and Jansen, 2002).

ACKNOWLEDGMENTS: The author is grateful to Dr. K. Chen, Mrs. Y. T. Li, Y. Bai, S. Z. Luo, T. Huang and Ms. X. F. Zhao for technical assistance.

ABBREVIATIONS: U, uracil; C, cytosine; A, adenine; G, guanine. aa, amino acid; A(Ala), P(Pro), V(Val), G(Gly), T(Thr), S(Ser), L(Leu), R(Arg), D(Asp), E(Glu), M(Met), I(Ile), F(Phe), C(Cys), W(Trp), H(His), Q(Gln), N(Asn), K(Lys) and Y(Tyr).

FOOTNOTES:

‡ Corresponding author: Fax: + 86 22 2350 3863. E-mail address: yangchm@nankai.edu.cn

REFERENCES

- Arques, D. G., Michel, C. J., 1996. A complementary circular code in the protein coding genes. *J. Theor. Biol.* 182, 45-58.
- Balakrishnan, J., 2002. Symmetry scheme for amino acid codons. *Phys. Rev. E. Stat. Nonlin. Soft Matter. Phys.* 65, 2 Pt 1.
- Bashford J. D., Tsohantjis, I., Jarvis, P. D., 1998. A supersymmetric model for the evolution of the genetic code. *Proc. Natl. Acad. Sci. USA* 95, 987-992.
- Bashford, J. D., Jarvis, P. D., 2000. The genetic code as a periodic table: algebraic aspects. *BioSystems* 57, 147-161.
- Bertman, M.O., Jungck, J.R., 1979. Group graph of the genetic code. *The Journal of Heredity* 70, 379-384.
- Blomberg, C., 1997. On the appearance of function and organisation in the origin of life. *J. Theor. Biol.* 187, 541-54.
- Cedergren, R., Miramontes, P., 1996. The puzzling origin of the genetic code. *Trends Biochem. Sci.* 21, 199-200. Erratum in: 1996. *Trends Biochem. Sci.* 21, 396. Comment in: 1997. *Trends Biochem. Sci.* 22, 49-50.
- Chipens, G. I., Gnilomedova, L. E., Ievinia, N. G., Kudriavtsev, O. E., Rudzish, R. V., 1989. The symmetry of the genetic code and the conservative nature of allowed mutations as a factor in evolution. *Zh. Evol. Biokhim. Fiziol.* 25, 654-63.
- Chipens, G. I., 1991. Asymmetry in the symmetric structure of the genetic code. *Zh. Evol. Biokhim. Fiziol.* 27, 522-9.
- Chipens, G. I., 1991. Hidden symmetry of the genetic code and laws of amino acid interaction. *Bioorg. Khim.* 17, 1335-46.
- Crick, F. H. C., 1968. The origin of the genetic code. *J. Mol. Biol.* 38, 367-379.
- Davydov, O. V., 1998. Amino acid contribution to the genetic code structure: end-atom chemical rules of doublet composition. *J. Theor. Biol.* 193, 679-690.
- Di Giulio, M., Capobianco, M. R., Medugno, M., 1994. On the optimization of the physicochemical distances between amino acids in the evolution of the genetic code. *J. Theor. Biol.* 168, 43-51.
- Di Giulio, M., Medugno, M., 1998. The historical factor: the biosynthetic relationships between amino acids and their physicochemical properties in the origin of the genetic code. *J. Mol. Evol.* 46, 615-621.
- Di Giulio, M., 2001. A blind empiricism against the coevolution theory of the origin of the genetic code. *J. Mol. Evol.* 53,724-732.
- Di Giulio, M., Medugno, M., 2001. The level and landscape of optimization in the origin of the genetic code. *J. Mol. Evol.* 52, 372-82.
- D'Onofrio, G., Jabbari, K., Musto, H., Bernardi, D., 1999. The correlation of protein hydropathy with the base composition of coding sequences. *Gene* 238, 3-14.
- Döring, V., Mootz, H. D., Nangle, L. A., Hendrickson, T. L., de Crécy-Lagard, V., Schimmel, P., Marlière, P., 2001. Enlarging the amino acid set of *Escherichia coli* by infiltration of the valine coding pathway. *Science* 292, 501-504.
- Dufton, M. J., 1997. Genetic synonym quotas and amino acid complexity: cutting the cost of proteins? *J. Theor. Biol.* 187, 165-173.
- Eigen, M., 1971. Self-organization of matter and the evolution of biological macromolecules. *Naturwiss* 58, 465-532.
- Eigen, M., Schuster, P., 1977. The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. *Naturwissenschaften* 64, 541-65.
- Eigen, M., Schuster, P., 1978. The hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle. *Naturwissenschaften* 65, 341-369.
- Eigen, M., Schuster, P., 1979. The hypercycle: a principle of natural self-organization. Springer-Verlag, Heidelberg.
- Eigen, M., Gardiner, W., Schuster, P., Winkler-Oswatitsch, R., 1981. The origin of genetic information. *Sci. Am.* 244, 88-92, 96, et passim
- Findley, G. L., Findley, A. M., McGlynn, S. P., 1982. Symmetry characteristics of the genetic code.

- Proc. Natl. Acad. Sci. USA 79, 7061-5.
- Frappat, L., Sorba, P., Sciarrino, A., 1999. Symmetry and codon usage correlations in the genetic code. *Phys. Lett. A* 259, 339-348.
- Frappat, L., Sciarrino, A., Sorba, P., 2001. Crystalizing the genetic code. *J. Biol. Phys.* 27, 1-34.
- Freeland, S.J., Knight, R.D., Landweber, L.F., Hurst, L.D., 2000. Early fixation of an optimal genetic code. *Mol. Biol. Evol.* 17, 511-8.
- Freeland, S. J., Hurst, L. D., 1998. The genetic code is one in a million. *J. Mol. Evol.* 47, 238-48.
- Gatlin, L. L., 1972. *Information Theory and the Living System: Ch. 6* (Columbia University Press, New York).
- Grantham, R., 1980. Working on the genetic code. *Trends Biochem. Sci.* 5, 327-333.
- Haig, D, Hurst, L. D., 1999. A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* 49, 708.
- Haig, D, Hurst, L. D., 1991. A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* 33, 412-7. Erratum in: 1999. *J. Mol. Evol.* 49, 708. Comment in: 1993. *J. Mol. Evol.* 49, 662-4.
- Halitsky, D., 1994. A geometric model for codon recognition logic. *Math. Biosci.* 121, 227-234.
- Hornos, J. E. M., Hornos, Y. M. M., 1993. Algebraic model for the evolution of the genetic code. *Phys. Rev. Lett.* 71, 4401 - 4404.
- Hornos, J. E. M., Hornos, Y. M. M., Forger, M., 1999. Symmetry and symmetry breaking: An algebraic approach to the genetic code. *Intl. J. Modern Phys. B* 13, 2795-2885.
- Jiménez-Montaño, M. A., de la Mora-Basañez, R., Pöschel, T., 1996. The hyperstructure of the genetic code explains conservative and non-conservative amino acid substitutions in vivo and in vitro. *BioSystems* 39, 117-125.
- Jiménez-Montaño, M. A., 1999. Protein evolution drives the evolution of the genetic code and vice versa. *BioSystems* 54, 47-64.
- Krakauer, D. C., Jansen, V. A. 2002. Red queen dynamics of protein translation. *J. Theor. Biol.* 218, 97-109.
- Kauffman, S.A., 1993. *The origins of order* (Oxford University Press, New York.)
- Knight, R. D., Landweber, L. F., 1998. Rhyme or reason: RNA-arginine interactions and the genetic code. *Chem. Biol.* 5, R215-20.
- Knight, R. D., Freeland, S. J., Landweber, L. F., 1999. Selection, history and chemistry: the three faces of the genetic code. *Trends Biochem. Sci.* 24, 241-7.
- Koch, A. J., Lehmann, J., 1997. About a symmetry of the genetic code. *J. Theor. Biol.* 189, 171-4.
- Karasev, V. A., Sorokin, S. G., 1997. The topological structure of the genetic code. *Genetika* 33, 744-51.
- Karasev, V. A., Stefanov, V. E. 2001. Topological nature of the genetic code. *J. Theor. Biol.* 209, 303-17.
- Lacan, J., Michel, C. J., 2001. Analysis of a circular code model. *J. Theor. Biol.* 213, 159-70.
- Lanzavecchia, S., Cantele, F., Radermacher, M., Bellon, P. L., 2002. Symmetry embedding in the reconstruction of macromolecular assemblies via the discrete Radon transform. *J. Struct. Biol.* 137, 259-72.
- Lehmann, J., 2000. Physico-chemical constraints connected with the coding properties of the genetic system. *J. Theor. Biol.* 202, 129-144.
- Liljenstrom, H., Blomberg, C., 1987. Site dependent time optimization of protein synthesis with special regard to accuracy. *J. Theor. Biol.* 129, 41-56.
- Luo, L. F., 1989. The distribution of amino acids in the genetic code. *Orig. Life Evol. Biosph.* 19, 621-31.
- Morphol. Embryol., Physiol., PHYSIOLOGIE* 22, 117-120.
- Portelli, C., Portelli, A. P., 1985. The symmetries of the genetic code of mammalian mitochondria.
- Racaniello, V. R., 1996. Early events in poliovirus infection: virus-receptor interactions. *Proc. Natl. Acad. Sci. U S A.* 21, 11378-81.
- Rakocevic, M., Jokic, A., 1996. Four stereochemical types of protein amino acids: synchronic determination with chemical characteristics, atom and nucleon number. *J. Theor. Biol.* 183, 345-9.
- Rodin, S., Ohno, S., Rodin, A., 1993. Transfer RNAs with complementary anticodons: could they reflect early evolution of discriminative genetic code adaptors? *Proc. Natl. Acad. Sci. USA* 90, 4723-4727.
- Rosen, G., 1999. Molecular content relations in the genetic code. *Phys. Lett. A* 253, 354-357.
- Shcherbak, V. I., 1988. The co-operative symmetry of the genetic code. *J. Theor. Biol.* 132, 121-4.
- Shcherbak, V. I., 1989. Rumer's rule and transformation in the context of the co-operative symmetry of the genetic code. *J. Theor. Biol.* 139, 271-6.

- Shcherbak, V. I., 1993. Twenty canonical amino acids of the genetic code: The arithmetical regularity. *J. Theor. Biol.* 162, 399-401.
- Siemion, I. Z., Stefanowicz, P., 1992. Periodical changes of amino acid reactivity within the genetic code. *BioSystems* 27, 77-84.
- Swanson, R. A., 1984. A unifying concept for the amino acid code. *Bull. Math. Biol.* 46, 187-203.
- Szathmary, E., 1999. The origin of the genetic code-amino acids as cofactors in an RNA world. *Trends Genet.* 15, 223-229.
- Trifonov, E. N., Bettecken, 1997. Sequence fossils, triplet expansion, and reconstruction of earliest codons. *Gene* 205, 1-6.
- Trifonov, E. N., 2000. Consensus temporal order of amino acids and evolution of the triplet code. *Gene* 261, 139-51
- Vargas, J. M., Stephens, C. R., Waelbroeck, H., Zertuche, F., 1999. Symmetry breaking and adaptation: evidence from a 'toy model' of a virus. *Biosystems* 51, 1-14.
- Weber, A. L., Lacey, J. C. Jr., 1978. Genetic code correlations: amino acids and their anticodon nucleotides. *J. Mol. Evol.* 11, 199-210.
- Weber, A. L., Miller, S. L., 1981. Reasons for the occurrence of the twenty coded protein amino acids. *J. Mol. Evol.* 17, 273-284.
- Wang, L., Brock, A., Herberich, B., Schultz, P. G., 2001. Expanding the genetic code of *Escherichia coli*. *Science* 292, 498-500
- Woese, C., 1967. The genetic code: The molecular basis for gene expression; Ch 6-7(pp. 156-160) (Harper and Row, New York/Evanston/London.).
- Woese, C. R., 1965. On the origin of the genetic code. *Proc. Natl. Acad. Sci. USA* 54, 1546-1552.
- Woese, C. R., Dugre, D. H., Saxinger, W. C., Dugre, S. A., 1966. The molecular basis for the genetic code. *Proc. Natl. Acad. Sci. USA* 55, 966-974.
- Wolpert, L., Lewis, J. H., 1975. Towards a theory of development. *Fed. Proc.* 34, 14-20.
- Wong, J. T., 1975. A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci. USA* 72, 1909-1912.
- Yarus, M., 2000. RNA-ligand chemistry: a testable source for the genetic code. *RNA* 6, 475-84.
- Yarus, M., 1998. Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin. *J. Mol. Evol.* 47, 109-117
- Yang, C. M. In a previous paper, <http://preprint.chemweb.com/biochem/0306001> (Date/Time of upload: 8 June 2003/09:21:34
- Ycas, M., 1999. Codons and hypercycles. *Orig. Life Evol. Biosph.* 29, 95-108.

(June 2003)