Available online at www.sciencedirect.com
SCIENCE DIRECT®
BioSystems 70 (2003) 187–209
BioSystems
www.elsevier.com/locate/biosystems

Arithmetic inside the universal genetic code

Vladimir I. shCherbak

Department of Applied Mathematics, al-Faraby Kazakh National University, 71 al-Faraby Avenue, Almaty 480078, Kazakhstan, CIS

Received 28 October 2001; received in revised form 23 April 2002; accepted 10 December 2002

Abstract

The first information system emerged on the earth as primordial version of the genetic code and genetic texts. The natural appearance of arithmetic power in such a linguistic milieu is theoretically possible and practical for producing information systems of extremely high efficiency. In this case, the arithmetic symbols should be incorporated into an alphabet, i.e. the genetic code. A number is the fundamental arithmetic symbol produced by the system of numeration. If the system of numeration were detected inside the genetic code, it would be natural to expect that its purpose is arithmetic calculation e.g., for the sake of control, safety, and precise alteration of the genetic texts. The nucleons of amino acids and the bases of nucleic acids seem most suitable for embodiments of digits. These assumptions were used for the analyzing the genetic code.

The compressed, life-size, and split representation of the *Escherichia coli* and *Euplotes octocarinatus* code versions were considered simultaneously. An exact equilibration of the nucleon sums of the amino acid standard blocks and/or side chains was found repeatedly within specified sets of the genetic code. Moreover, the digital notations of the balanced sums acquired, in *decimal* representation, the unique form 111, 222, . . . , 999. This form is a consequence of the criterion of divisibility by 037. The criterion could simplify some computing mechanism of a cell if any and facilitate its computational procedure. The cooperative symmetry of the genetic code demonstrates that possibly a zero was invented and used by this mechanism. Such organization of the genetic code could be explained by activities of some hypothetical molecular organelles working as natural biocomputers of digital genetic texts.

It is well known that if mutation replaces an amino acid, the change of hydrophobicity is generally weak, while that of size is strong. The antisymmetrical correlation between the amino acid size and the degeneracy number is known as well. It is shown that these and some other familiar properties may be a physicochemical effect of arithmetic inside the genetic code.

The “frozen accident” model, giving unlimited freedom to the mapping function, could optimally support the appearance of both arithmetic symbols and physicochemical protection inside the genetic code.

Keywords: Genetic code; Genetic language; Arithmetic power; Natural biocomputer; Mathematical model

1. Introduction

Arithmetic is the only tool for producing information systems of extremely high efficiency. Could life, being an information phenomenon, use arithmetic for the sake of control, safety, and precise alterations of its genetic texts? To answer this question, we can resort to a similarity among genetic, ordinary, and algorithmic languages (Eigen and Winkler, 1985).

Ordinary languages fuse the letters and digits into a common alphabet of signs. A meaning is mapped onto these signs by virtue of linguistic condition of abstract character (Yockey, 2000). For digits, this condition contains the numeration system. The only reason for the numeration system to appear is for arithmetic calculations. Hence, the presence of the

E-mail address: genecodelab@hotmail.com (V.I. shCherbak).

0303-2647/$ – see front matter
doi:10.1016/S0303-2647(03)00066-2

numeration system inside the alphabet of some language can be an indication that users of that language know and employ arithmetic.

The genetic code is a *sui generis* alphabet of genetic texts. The digital nature of these texts was noted for various reasons (Gamow, 1954; Yockey, 1992, 2000), however, computing has not been noted in genetic text. As regards to amino acids, their digital nature is not clear. Hydrophobicity—an important property of the amino acids—can hardly be digitized (Jiménez-Montaño, 1993; Siemion, 1994). That is why we have to discern a more fine and invariable embodiment of a digit in amino acids. Thus, the nucleons of amino acids, i.e. an equivalent of their mass, and the bases of genetic supermolecules (Gamow, 1954) seem more suitable for "balls" of a biological "abacus".

An arithmetic power in the linguistic milieu of genetics is possible. The *artificial* biocomputers working by using the genetic supermolecules were discussed theoretically and tested experimentally (Adleman, 1994; Chen and Wood, 2000; Faulhammer et al., 2000). Continuing such ideas, one may assume that *natural* computing can exist as well. For instance, such a computing could be essence of exact gene processing and scrambling (Landweber and Gilbert, 1993; Landweber et al., 2000). If that is the case, some cell organelles should work as biocomputers. Thereby, we have to discover the number systems with which they work. We can imagine that at the moment of the origin of the arithmetic power, the biocomputers modified the genetic code itself. Indeed, any given alphabet and the genetic code in particular should somehow correspond to the selected number system. Moreover, the ordinary computers suggest that some useful feature of the number system can totally predetermine genetic "soft hardware".

The number system is not a physicochemical property and, hence, only our mind can discern this arithmetic abstract. It was noted that the genetic code has much in common with the ASCII code (Yockey, 2000), which is computer alphabet of algorithmic languages. Binary-numerated and sequenced ASCII symbols reproduce, in turn, the natural series of the Arabic numerals, starting with a zero, as well as alphabetically, i.e. semantically, arranged Roman letters. In that way, the linguistic essence as well as internal binary and decimal arithmetic of the ASCII code become apparent.

Could the similarly arranged genetic code have analogous regulation? Specifying some relevant code sets, Gamow (1954), Rumer (1966), and Hasegawa and Miyata (1980) did not suppose that some elements of these sets emphasize the decimal notation of the positional number system. Moreover, remarkable physical relations support this arithmetic abstract. As shown below, "decimalization" is superposed perfectly onto the numerous exact balances of the nucleons as well as onto the semantic symmetry of the triplets, a zero also acts as a special number.

When we think of the genetic code, we usually visualize certain physicochemical structures and their interactions (Knight et al., 1999) or ask for some physical relations among them (Sukhodolets, 1989; Rosen, 1999). Although these are burning issues, it seems that the genetic code is connected more closely to abstract notions of arithmetic than with notions of physics or chemistry. Let us give detailed reasons for that.

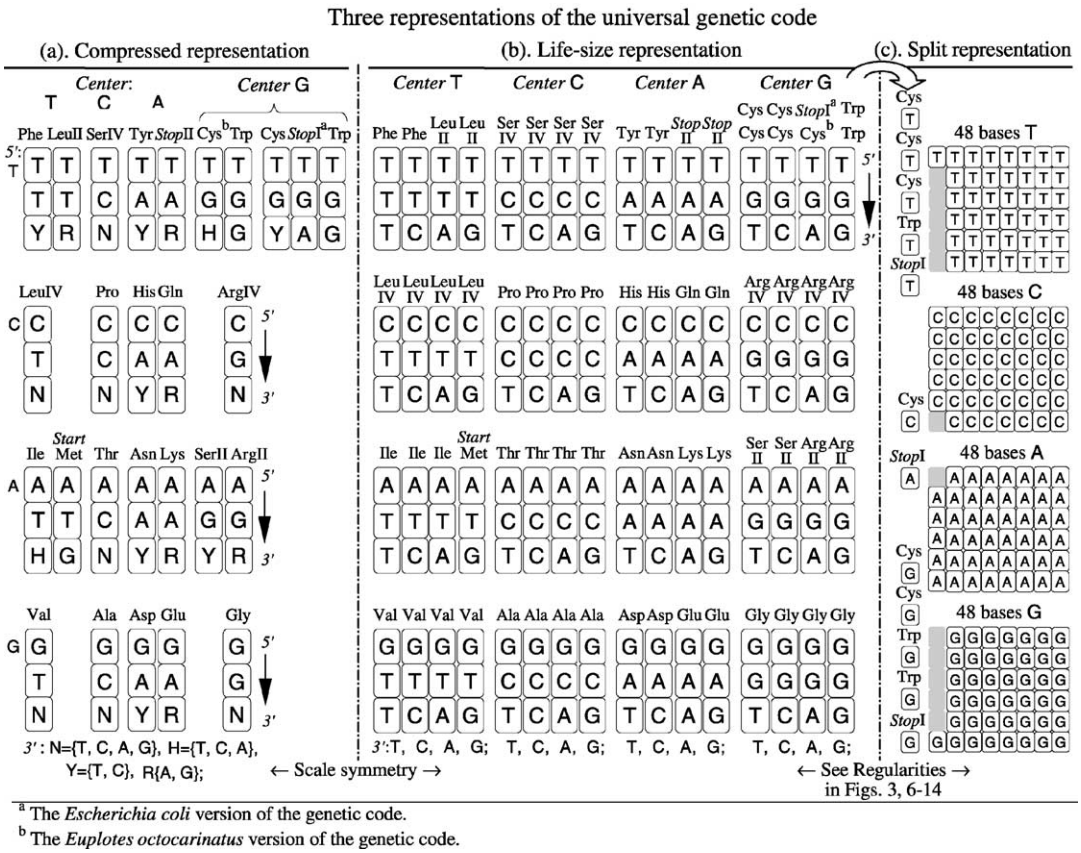
2. A few relevant images of the genetic code

The universal genetic code is shown in Fig. 1. From the mathematical viewpoint, it is the mapping function $f{:}T \rightarrow A,S$ of the 64 base triplets (T) and 20 L-α-amino acids (A) with two syntactic signs *Start* and *Stop* (S). The *compressed* representation of the genetic code is formed by the series of synonymic triplets whose 3′ bases are compressed on the 1:D scale; where D is the degeneracy numbers IV, III, II, I (Fig. 1a).

The *life-size* representation includes 64 pairs of the *triplet-amino acid* and *triplet-syntactic sign* type (Fig. 1b). The regular degeneracy and universal triplet are the general ordering of the genetic code, which have not had a plausible explanation until now.

The *split* representation (Fig. 1c) completes antisymmetrically its compressed form. The split representation is made up of four base types with 48 identical molecules of each type. The simple rule maps 192 separate bases onto 20 amino acids and 2 syntactic signs. The separate bases can be transformed inversely into triplets by the conditions specified in Figs. 3 and 6–14.

Let us consider two versions of the genetic code. These are the commonly accepted *Escherichia coli*



[a] The *Escherichia coli* version of the genetic code.
[b] The *Euplotes octocarinatus* version of the genetic code.

Fig. 1. The universal genetic code. There are the *E. coli* code version with the TGY-Cys, TGA-*Stop*I series and the *E. octocarinatus* code version with the TGH-Cys series. (a) The compressed representation of the genetic code includes the synonymic series whose 3′ bases are compressed mathematically to the 1:*D* scale, where *D* is the degeneracy number. (b) The life-size genetic code representation includes 64 triplets, onto which 20 amino acids and 2 syntactic signs, *Start* and *Stop*, are mapped. (c) The split genetic code representation places four types of the nitrogen bases outside their triplets. The separate bases are combined into four sets. The amino acids and syntactic signs are mapped onto these bases by the rule that is demonstrated for the 5′ TG fragment of the *E. coli* version of the genetic code. The direction from the 5′ to 3′ position coincides with the top-to-bottom direction (this manner of writing allows some new code symmetries to be shown graphically).

version with the TGY-Cys, TGA-*Stop*I series and the *Euplotes octocarinatus* version with the TGH-Cys series (Marshal et al., 1967; Meyer et al., 1991; Tan and Heckmann, 1998; Grimm et al., 1998; Tan et al., 1999). Only the canonical 20 amino acids are studied. Their ionized forms, post-translational modifications, as well as the cotranslational incorporation of the SeCys amino acid are beyond the scope of this work.

Rumer (1966) partitioned the six time degenerated series into the degeneracy II and IV series and gathered the "complete" series of degeneracy IV into one set and the "broken" series into the other (Fig. 2a). Note that each of these sets gets the same number of the life-size triplets, which are mapped in a *one-to-one* manner by the unique Rumer's TCAG → GACT transformation or by combined TCAG → CTGA and TCAG → AGTC transformation (Fig. 3; Shcherbak, 1989a; Karasev and Sorokin, 1997). These transformations form an algebraically closed Abelian group of order 4, denoted below by V. Although each of these transformations can act independently or be absent, they constitute an ordered assembly. Dankwerts and Neubert (1975), Bertman and Jungck (1979),

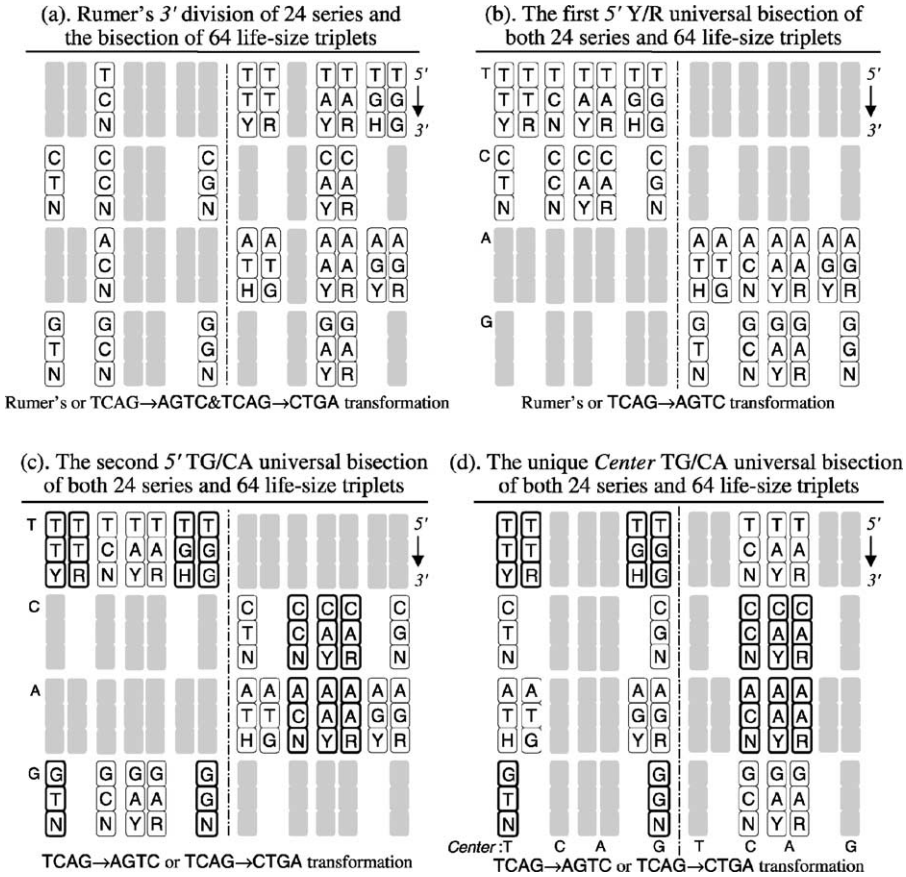


Fig. 2. The Rumer's (1966) division of the compressed genetic code (a) and three bisections (b), (c), and (d) suggested by Konopelchenko and Rumer (1975). Wide bordering shows the common series of the bisection (c) and (d) for the sake of regularities in Fig. 8. The *E. octocarinatus* code version is used for the representation of the compressed genetic code here and in the figures below because the degeneracy is distributed symmetrically among its series.

Jiménez-Montaño et al. (1996), Zhang (1997), and Jiménez-Montaño (1999) considered the group structure of the genetic code in this aspect.

Konopelchenko and Rumer (1975) noted other *universal* bisections of the compressed code. The universality means the simultaneous bisection of the 64 life-size triplets, 8 "complete", and 16 "broken" series, as well as the correspondence of the bisection axes to the transformations of group V. Due to natural distribution of the degeneracy, there are only two such 5′ bisections (Fig. 2b and c) and the unique *Center* bisection (Fig. 2d).

Gamow (1954) suggested one significant division of the 64 life-size triplets. He divided the triplets depending on the composition of identical and unique bases, regardless of their types and positions in the triplets (Fig. 4).

3. The number system is attribute of computing

If the number system were in the sets specified above, it could be disclosed by means of some feature of natural numbers produced by the system itself. The consequences of Pascal's general criterion of divisibility are but few of such features. Ordinary computers use the binary notation and even parity, i.e. divisibility by 2, as an accuracy control. Similarly, a cell could use the perfect positional system with enough small radix and some criterion of divisibility by the number

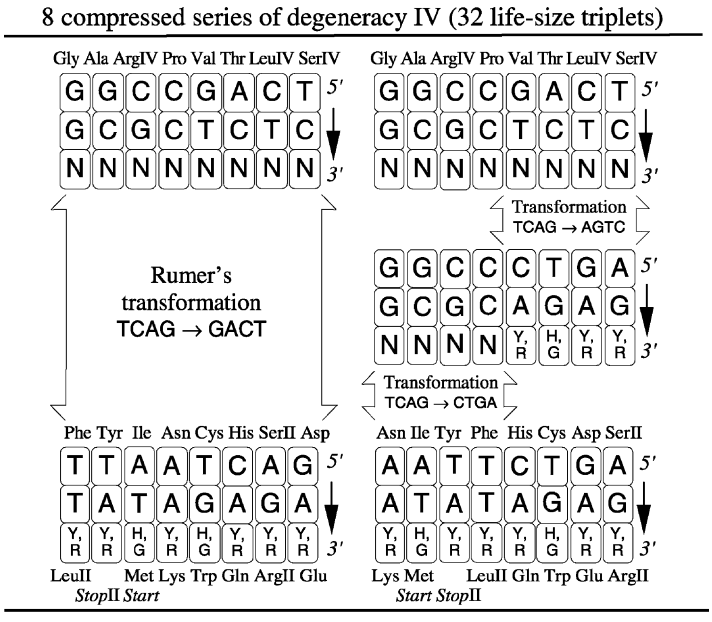


Fig. 3. The Rumer's (1966) division. The series of degeneracy IV and III, II, I are distributed among the two sets in the 1:2 ratios whereas the life-size triplets are bisected. The triplets are mapped either by the Rumer's unique transformation or by the combination of two other transformations of group V. The graphic representation of the genetic code images its *E. octocarinatus* version; the revealed regularities are also true for the *E. coli* version.

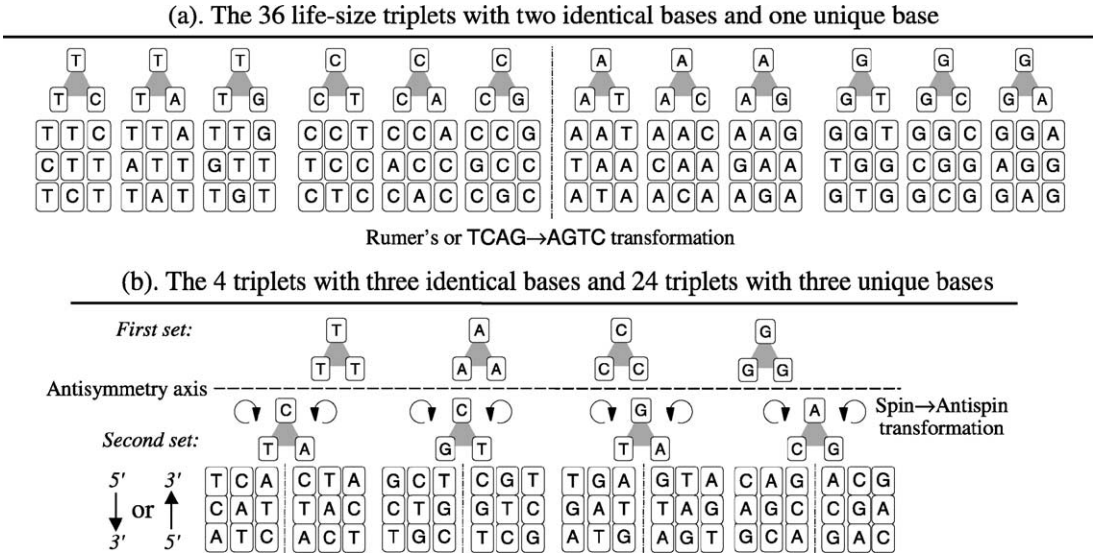


Fig. 4. The Gamow's division of the 64 life-size triplets. The triplets are divided into three sets depending on the base composition of the triplets. Twenty possible combinations of four bases, three at a time, are placed on triangular substrates. Each of the combinations generates some fixed triplets. (a) Triplets with two identical bases and one unique base form the largest set. (b) As regards to the other two sets, the first one comprises the identical base triplets, while the second one—antisymmetrically—comprises the unique base triplets. Pairs of circular arrows show *Spin* → *Antispin* transformation.

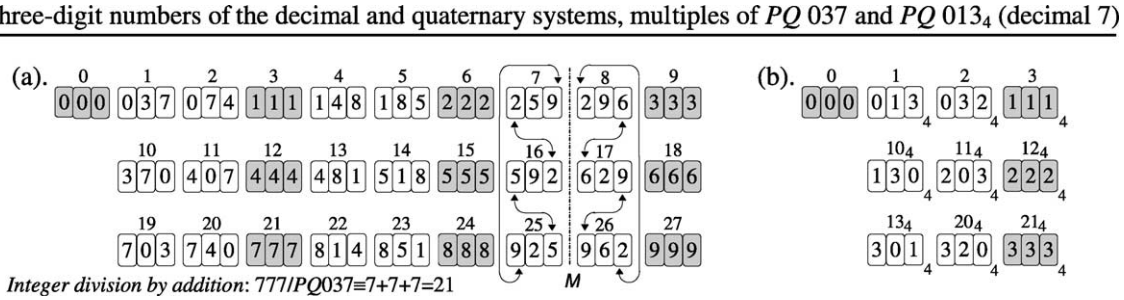


Fig. 5. Consequences of Pascal's general criterion of divisibility for three-digit numbers of the decimal and quaternary system that are divided by the PQ 037 and PQ $013_4 = 7$, respectively. Digital sequences of these numbers are ordered identically in both systems. Digital symmetries are exhibited with a period of $(q-1)$, where q is the radix of the number system. Numbers made up of the same digits have a period equal to their order $n = 3$. The sum of the digits within their notation is equal to the quotient of these numbers divided by PQ. Such sum for numbers made up of different digits is equal to the central quotient in the column. This feature allows one to organize the operation of integer division by PQ by the operation of addition. The numbers made up of different digits are linked by a cyclic permutation in the columns. The direction of permutation in the neighboring columns is mirror symmetrical; the letter *M* and the respective axis designate this symmetry. The digital regularities are invariant under the inversion of the direction of number notation. The notation of the numbers 037 and 013_4 with a leading zero emphasizes the equal opportunities for zero to participate in the digital symmetries. The present criterion of divisibility bears a close analogy to the formal parameters of the genetic encoding.

equal to 2 (Mac Dynaill, 2002) and/or larger than 2 to increase the accuracy of the control. This criterion should give particular shape to the digital sequence so that the molecular genetic machinery can recognize it easily. For example, the decimal criterion of divisibility by *Prime Quantum* (PQ) 037 fulfills this condition (Fig. 5a).

The criterion says: only the decimal numbers, whose three-digit sum of digital triplets has a homogeneous digital sequence (or the sum acquires such a sequence after adding/subtracting PQ 037), are divided by PQ 037 (the triplet frame has any of the three possible positions with zeros added at the flanks to form complete triplets). Let us consider this criterion for the number 60150013320229856, for example. The sum of the digital triplets for one of the frames is equal to the number $601 + 500 + 133 + 202 + 298 + 560 = 2294$. In order to get the three-digit sum the operation is repeated $002 + 294 = 296$ and, finally, PQ 037 is added to this sum $296 + 037 = 333$. The operations in other two frames are the same:

$060 + 150 + 013 + 320 + 229 + 856 = 1628,$
$001 + 628 = 629,\ 629 + \text{PQ}\,037 = 666;$
$006 + 015 + 001 + 332 + 022 + 985 + 600 = 1961,$
$001 + 961 = 962,\ 962 + \text{PQ}\,037 = 999.$

Only one such summation in any of the frame is sufficient for finding out the divisibility by PQ 037. Note that the digital sequences 296, 629, 962 are connected by a cyclic permutation of the digits, which is another "palpable" criterion feature. The criterion is valid for the PQ $= \langle 1_n 1_{n-1} \ldots 1_1 \rangle_q / n$, if the condition $(q-1)/n = \text{Int}$ is applied and the n-digit reading frame is used. For example, the systems with radixes $q = 4$ (PQ 7, Fig. 5b), $q = 7$ (PQ 19), $q = 10$ (PQ 37, Fig. 5a), $q = 13$ (PQ 61), etc. meet all these requirements for $n = 3$.

The symmetry of the three-digit sequences in Fig. 5 is named below as the *Prime Quantum Divisibility* (PQD) feature. It gets the particular shape of digital triplet if the decimal (quaternary) system is used and loses it in the other systems. The PQD feature not only excites the sensation of beauty but also could be useful for some molecular organelles working as biocomputers. This feature may simplify molecular machinery and facilitates the computational procedure itself. Such ability is discussed in Section 13.

4. The nucleons and nitrogen bases as digits

The number system might use a quantity of some code elements to symbolize a number. Such remarkable elements could come to light indirectly in one of the familiar code regularities and, most probably, in some correlation with degeneracy. Using

Table 1
The antisymmetrical correlation between the amino acid side chain nucleon number R and the series degeneracy number D (Hasegawa and Miyata, 1980)

Amino acid	Gly	Ala	Ser	Pro[a]	Val	Thr	Cys	Leu	Ile	Asn	Asp	Gln	Lys	Glu	Met	His	Phe	Arg	Tyr	Trp
Side chain nucleon number, R	1	15	31	42–1	43	45	47	57	57	58	59	72	72	73	75	81	91	100	107	130
Degeneracy number, D	IV	IV	IV, II	IV	IV	IV	III[b], II[c]	IV, II	III	II	II	II	II	II	I	II	II	II, IV	II	I

The wedge-shaped figure and the line of variable thickness denote a change of the nucleon number and the degeneracy number, respectively. This correlation is a part of the genetic code cooperative symmetry in Fig. 12.

[a] Formal borrowing of one nucleon from the Pro side chain in favor of its standard block (see text for details).

[b] The *E. octocarinatus* code version.

[c] The *E. coli* code version.

degeneracy—a mutual product of both code sets—Rumer ordered the set of triplets; for symmetry, the degeneracy should order the amino acid-syntactic sign set as well. Indeed, Hasegawa and Miyata (1980) noted that the degeneracy number D and the amino acid nucleon number N are antisymmetrically correlated (Table 1). Let us use the nucleon as the embodiment of a digit for the amino acids. The most common and stable isotopes are taken for the calculating the nucleon numbers of the 20 canonical amino acids.

Each of the amino acid molecules is divided universally into two parts. These are the unique *side chain R* and the *standard block B* –CH(NH_2)COOH. For instance, the nucleon number of the least Gly and the largest Trp side chain equals 1 and 130. The nucleon number of the syntactic signs *Stop*I and *Stop*II, which do not have corresponding amino acids, is set to be zero; the syntactic sign *Start* associates with the amino acid Met whose nucleon number is 75, etc. The one exception to the general structure of amino acids is Pro. It is called an *imino* acid because the nitrogen of its standard block has one less hydrogen bonded to it. In the special case of Pro, the division produces a block, which has 73 nucleons. The formal borrowing of one nucleon from the Pro side chain in favor of its block brought the block nucleon number to the standard $73 + 1 = 074$, whereas the side chain nucleon number becomes $42 - 1 = 41$. Seemingly strange from the physicochemical point of view, the borrowing looks like an ordinary computational trick of hypothetical biocomputers.

Another number symbol may be composed by the rules of the positional system of numeration. The nitrogen bases of the genetic supermolecules seem suitable to the role of the quaternary digits. Gamow (1954) noted this possibility in his pioneering article on the genetic code. He wrote that a long number written in a four-digital system could characterize the hereditary properties of any given organism. Time has shown that his statement was taken as a metaphoric one.

5. The arithmetic regularity of the compressed code representation

Modified by biocomputers, the genetic code should add a certain formal ordering to its well-known physicochemical regularities. Let us consider the four levels of this ordering. The *logic* level includes the universal conditions by which the code is divided into sets. The cardinal numbers of these sets are studied at the *physical* level to find some quantitative relations, whereas the digital sequences of these numbers are studied at the *linguistic* level to find the PQD features. And finally, a cooperative behavior of the code elements is studied at the level of the *ordered* sets.

Rumer discovered the ordered manifestation within the triplet set (Fig. 3). The supplementary manifestation within the amino acid-syntactic sign set is shown in Fig. 6 (Shcherbak, 1993b, 1994). The logic level contains the condition of Rumer's division. The physical and linguistic levels are superimposed exactly in spite of their independence. The physical level is manifested by the squares of the first three Pythagorean numbers (Fig. 6a); these numbers appear if the nucleon sums are represented in the parts of PQ 037. Another manifestation of the physical level is a balance of the nucleon sums of the standard blocks and side chains

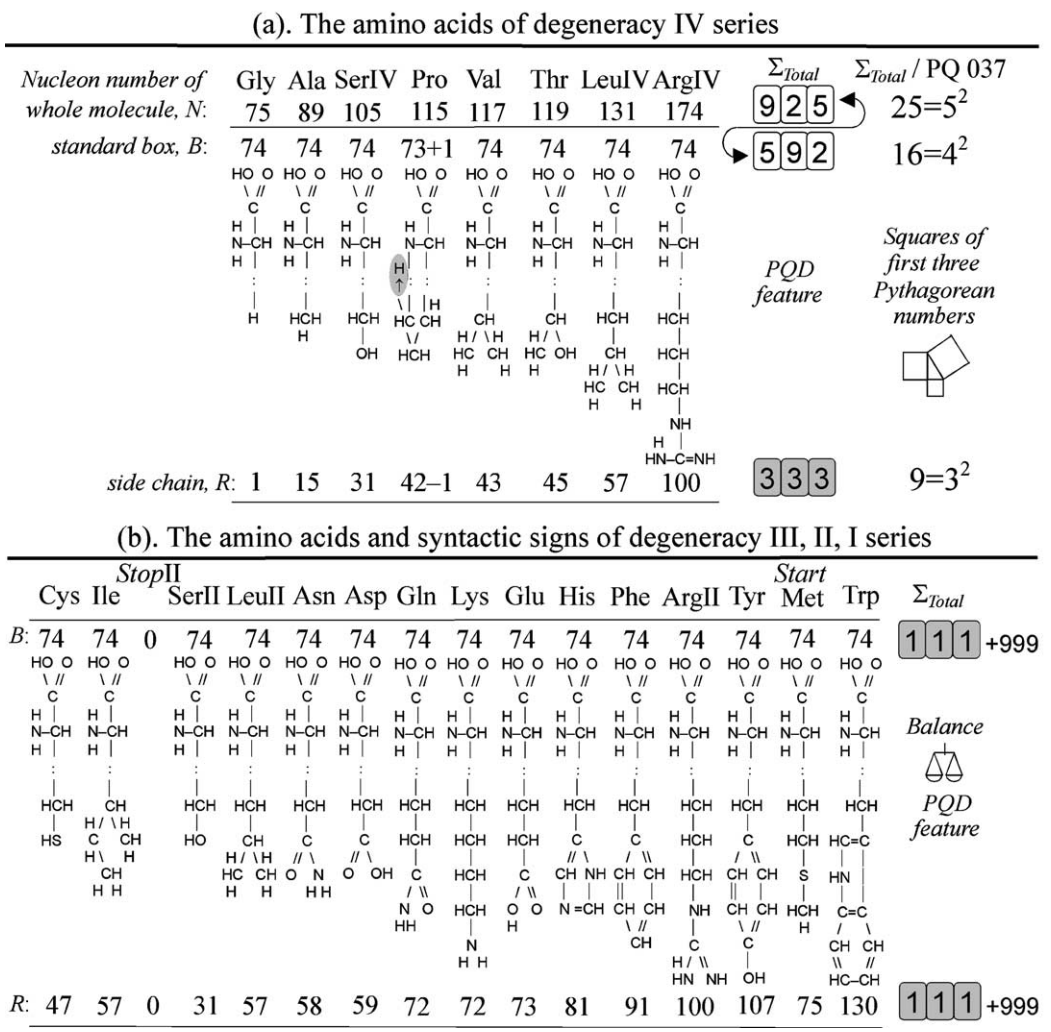


Fig. 6. The arithmetic regularity of the compressed representation of the genetic code (see Figs. 2a and 3). It appears simultaneously at the logic, physical, and linguistic level. The logic level contains the condition of the Rumer's division. The physical level shows the quantitative properties of the nucleon sums. These are (a) the squares of the first three Pythagorean numbers and (b) the *block-and-chain* balance. The linguistic level manifests the PQD feature. The raster spot indicates the formal borrowing of one nucleon from imino acid Pro side chain in favor of its standard block. Numbers, divided by PQ 037 and larger 999, are written ad exemplum $2331 \equiv 333 + 2 \times 999$ to underline their PQD features. Note that the product of PQ 037 and PQ 013_4 is equal to 259, which is the final digital permutation in (a). The canonical DNA pair of the adenine and thymine residues has also 259 + "0" nucleons and the other pair of the guanine and cytosine residues has 259 + "1" nucleons; the alone uracile residue has 111 nucleons. The graphic representation of the genetic code images its *E. octocarinatus* version; the revealed regularities are also true for the *E. coli* version.

in Fig. 6b. The linguistic level represents the PQD 037 features of the nucleon sums. Such a multilevel manifestation is referred to as an *arithmetic* regularity.

The nucleon sum of the standard blocks of the whole code is larger than that of the side chains. Chains are larger than blocks only in 7 out of 24 series. Only a limited set with the number of series less than 24 and containing mostly large amino acids is capable of balancing their own blocks. Owing to requirements of Eq. (13) (Table 2) the balance together with large amino acids is located within the degeneracy III, II, I set and not within the IV set of the genetic code.

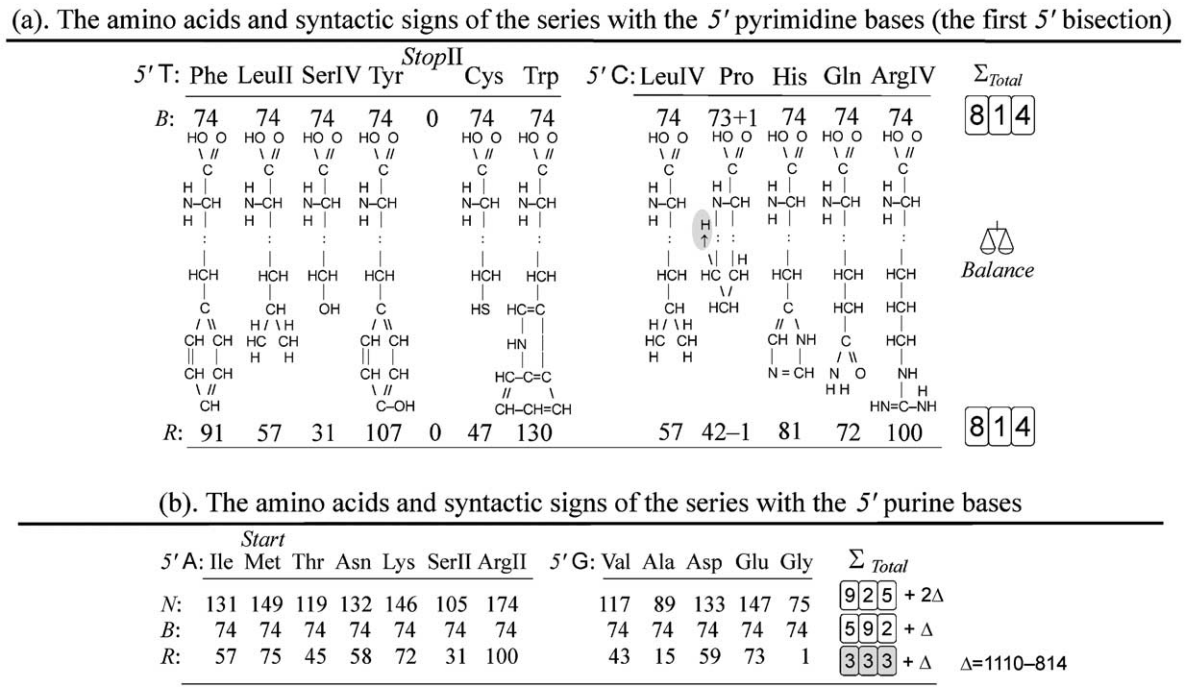


Fig. 7. The arithmetic regularity of the compressed representation of the genetic code (see Fig. 2b). (a) The amino acids and syntactic signs whose triplet series have the pyrimidine base in the 5′ triplet position manifest a new block-and-chain balance. (b) The arithmetic regularity in Fig. 6 and the new balance algebraically predetermine the quantitative relations. The graphic representation of the genetic code images its *E. octocarinatus* version; the revealed regularities are also true for the *E. coli* version.

As a result, the concentration of small amino acids increases in the degeneracy IV set. Outwardly, this looks like antisymmetrical correlation between the degeneracy and molecular mass noted by Hasegawa and Miyata (1980). The regularities in Fig. 6 are a possible interpretation of this correlation.

A similar response in analogous circumstances indicates a general regularity of the investigated object. The amino acids of the first 5′ bisection (Fig. 3b) whose series have 5′ pyrimidine bases are shown in Fig. 7a. The same elements, arranged differently, demonstrate a new *block-and-chain* balance, similar to the previous one in Fig. 6b. After the balance in Fig. 6b the PQD 037 feature becomes trivial in an algebraic sense for the block-and-chain type of balancing. Surprisingly, the PQD 037 feature changes henceforth the type of manifestation to bypass this limitation.

One manifestation of the arithmetic regularity makes use of the second 5′ and unique *Center* bisection simultaneously (Fig. 2c and d). These bisections comprise a hierarchy of three balances of the *chain-and-chain* type (Fig. 8). The ordering procedure in Figs. 7 and 8 were patterned by Verkhovod (1994) after it was discovered in Fig. 6 (Shcherbak, 1993a,b).

Surprisingly, the borrowing of one nucleon from the imino acid Pro side chain works, while the standard block itself being the recipient of such a borrowed nucleon, is located outside the considered manifestation.

Mitochondria deviate from the universal code. It should be noted that none of the numerous deviations (Jukes and Osawa, 1990) disturbs the balance of common series in Fig. 8c.

6. The arithmetic regularity of the life-size code representation

The triplets having two identical and one unique base are shown in Fig. 9 (see also Fig. 4a; Shcherbak, 1996). The arithmetic regularity is manifested within two bisected subsets whose triplets have either two identical pyrimidine or two identical purine bases

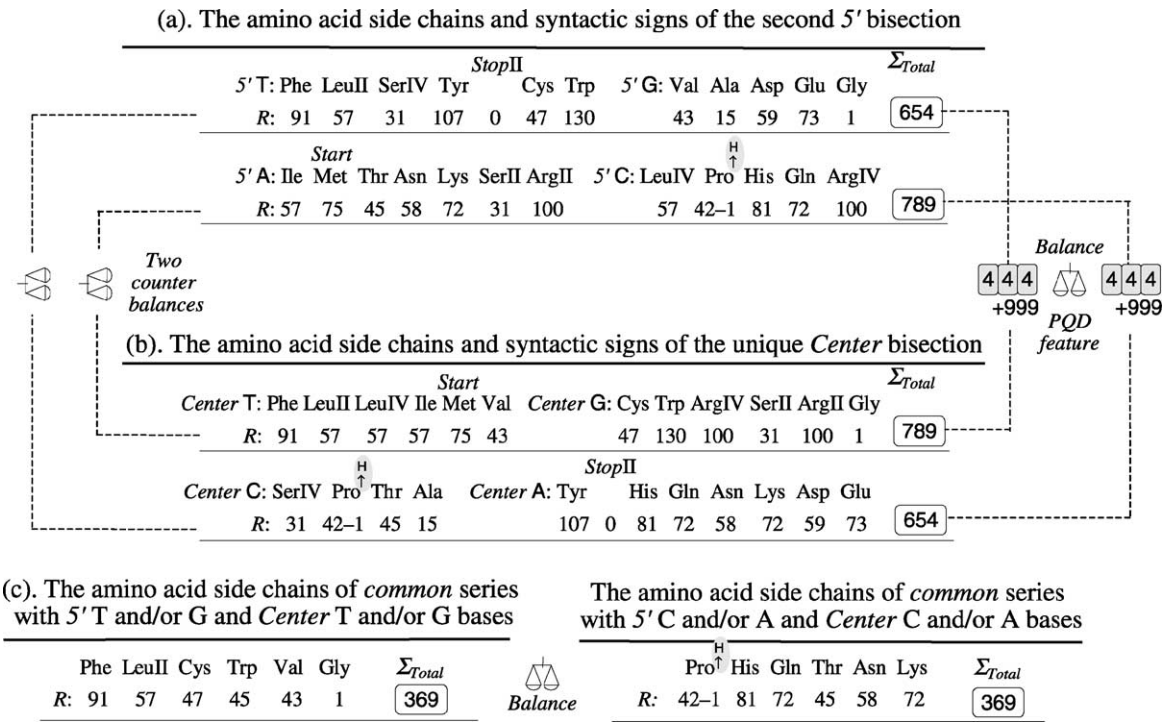


Fig. 8. The arithmetic regularity of the compressed representation of the genetic code (see Fig. 2c and d). The side chain nucleon sum of the series with 5′ and *Center* bases T, G is equal to the chain nucleon number 444 + 999 of the whole compressed code. The same is also true for the series with bases A, C (see right side of (a) and (b)). These *chain-and-chain* balances are caused by the counterbalance of 5′ T, G and *Center* A, C arms as well as the 5′ A, C and *Center* T, G arms (see left side of (a) and (b)). The equilibrated hierarchy as a whole is based on the common series balance (see (c)); wide bordering shows these series in Fig. 2c and d. The common series have T and/or G, and A and/or C bases in both 5′ and *Center* position. The graphic representation of the genetic code images its *E. octocarinatus* version; the revealed regularities are also true for the *E. coli* version.

(Fig. 9a). A new chain-and-chain balance is formed by 999 nucleons. The PQD feature is always significant for this type of balancing; the nucleon borrowing of the Pro amino acid works as well.

It is believed that the 3′ base is of less importance for the biological encoding. It seems that the 5′, *Center*, and 3′ bases are of equal importance for the arithmetic regularity.

In Fig. 9a the central axis bisects the code in such a way that the type of unique base is disregarded. Now, the central axis in Fig. 9b bisects the code by using the type of unique base, disregarding identical ones. The sum of nucleons for the triplets with the unique pyrimidine base is equal to the decimal number 888 and the PQD 037 feature is manifested again.

The subset with purines as identical bases in Fig. 9a has an additional manifestation of the arithmetic regularity. This subset consists of three lines of the triplets, which are subdivided symmetrically. Two of the subdivisions are a pair of halved lines (a vertical dash-line indicates a halving) and the third one is a complete line (shown by wide bordering). These subdivisions demonstrate a *triple* balance of the chain-and-chain type. Each of its balanced arms consists of 333 nucleons.

In addition, the complete line has an analogue written by synonymic triplets in Fig. 9b. This balanced pair of lines owes its existence to the regular degeneracy and, in particular, the nonordinary degeneracy III of the amino acid Ile. This is the reason why the sum of nucleons in Fig. 9b is given by two summands 333 + 777. That sum can be transformed into balance of the (*block* + *chain*)-*and-chain* type, i.e. 777 + 777.

The arithmetic regularity of the life-size code becomes complete with similar arrangement of the rest of

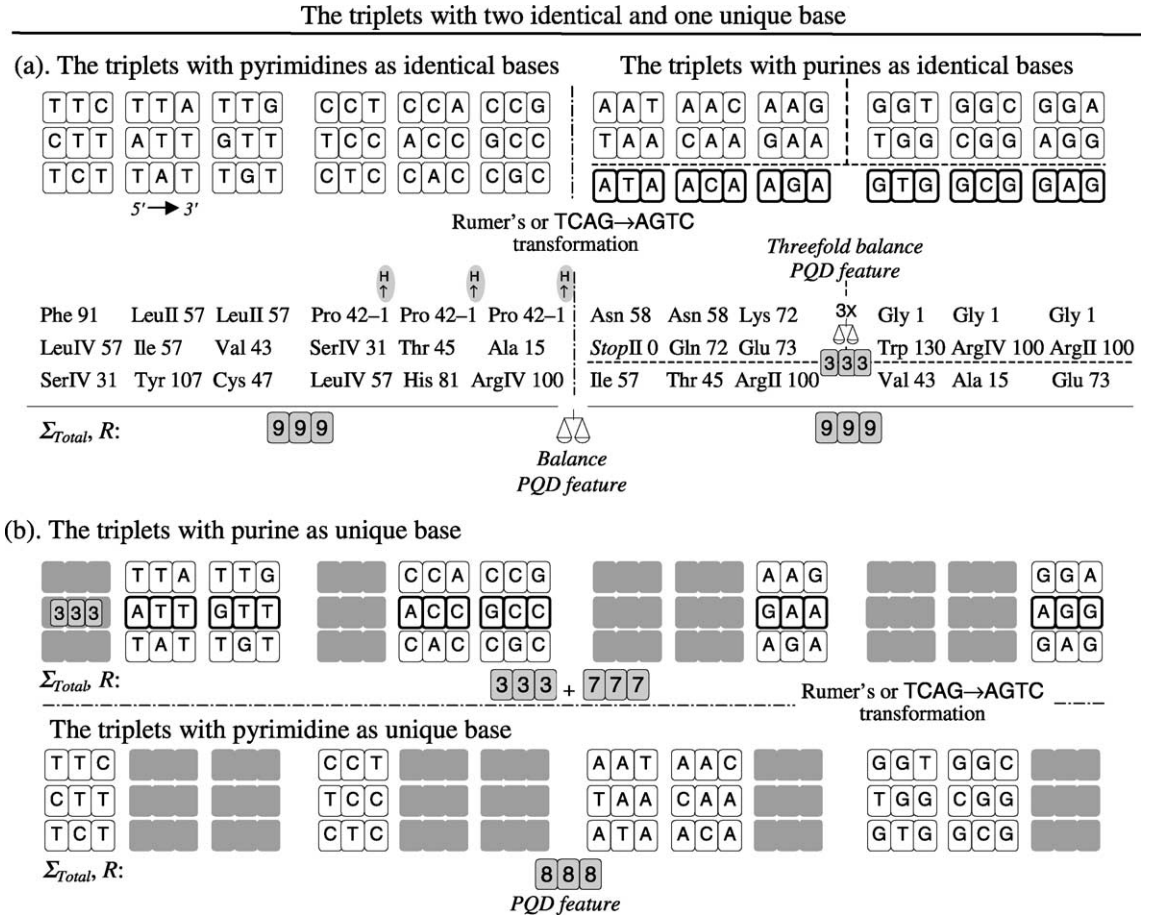


Fig. 9. The arithmetic regularity of the genetic code life-size representation (see Fig. 4a). (a) The triplets with the identical pyrimidine and identical purine bases manifest the chain-and-chain balance and PQD feature. The triplets with identical purine bases manifest the triple chain-and-chain balance and the PQD features. (b) The triplets with the unique pyrimidine and unique purine base manifest the PQD feature. (a) and (b): The synonymic triplets of two balanced lines are denoted by the wide bordering. The graphic representation of the genetic code images both the *E. coli* and the *E. octocarinatus* version.

the triplets. These are the first and second antisymmetrically arranged code sets in Fig. 5b. The antisymmetry underlies their (*block*+*chain*)-*and*-(*block*+*chain*) balance in Fig. 10 as well.

7. The arithmetic regularity of the split code representation

The split representation of the genetic code in Fig. 1c consists of $64 \times 3 = 192$ separated bases. The amino acids and syntactic signs are mapped onto these bases and listed in Fig. 11. The only division into the separate *Thymine*-list and the common *Adenine-Guanine-Cytosine*-list demonstrates the PQD 037 feature of the side chain nucleon sums $666 + 2 \times 999$ and $555 + 7 \times 999$, respectively.

As a rule, the linguistic level of the arithmetic regularity is superimposed on some balance, for example, of the block-and-chain type. Recall that chains all over the code may balance only some part of their blocks. The common *A-G-C*-list contains 138 standard blocks, whose nucleon sum is equal to $222 + 10 \times 999$. The nucleon sum of the amino acid side chains of both lists, i.e. the split code on the whole, is exactly the same number (Shcherbak, 1999).

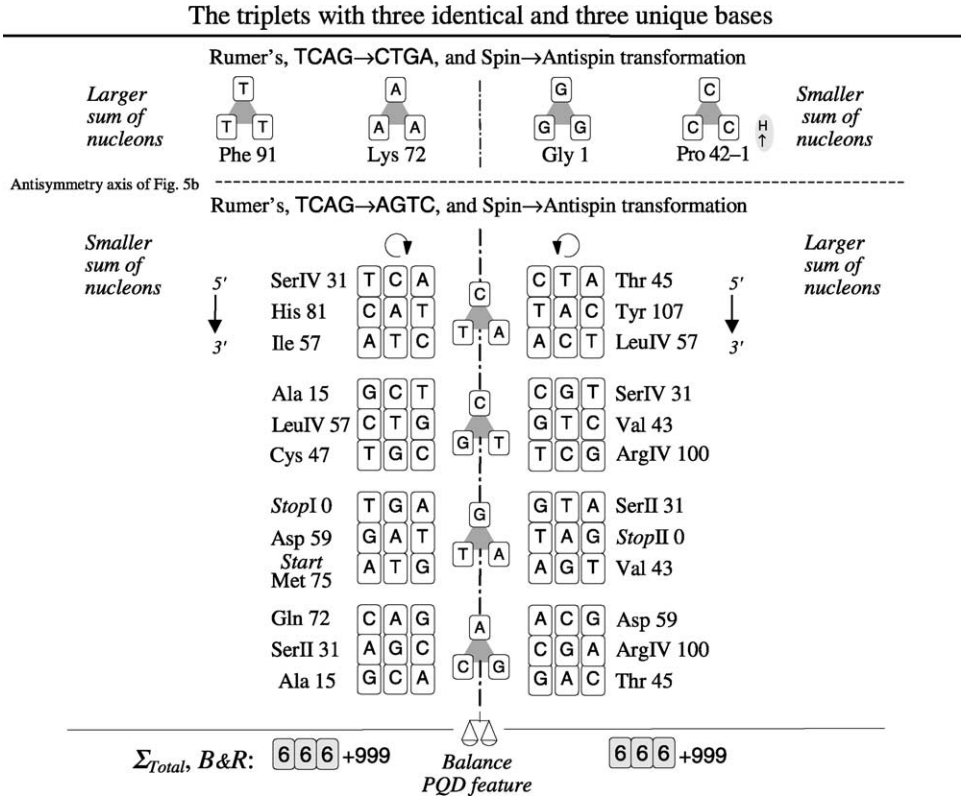


Fig. 10. The arithmetic regularity of the genetic code life-size representation (see Fig. 4b). There is a unique way of antisymmetrical application of the three group V transformations to the pair of the code sets. Namely, one of the transformations is chosen to be common for both sets; if in the first set it is paired with one of the remaining transformations, then in the second set it is paired with the other one. The *Spin* → *Antispin* transformation does not affect the triplets in the first set, but—antisymmetrically—it fixes the triplet positions in the second set. The sets could be coaxially placed in one of the alternative positions. The antisymmetrically positioned smaller and larger nucleon sums eliminate this alternative. The balance becomes apparent if the common Rumer's transformation is paired with the TCAG → CTGA transformation in the first set. The graphic representation of the genetic code images its *E. coli* version; the revealed regularities are true for this version.

New block-and-chain balance is similar to the previous balances in Figs. 6b and 7a, but is contrarwise organized. There, the fixed set of the chains balances all their own blocks, whereas here, the fixed set of the blocks balances all the code chains. The PQD feature becomes again a significant manifestation due to another block-and-chain type of balancing.

8. The systematization principle of the cooperative symmetry of the genetic code

The arithmetic regularity operates with unordered sets. The following is an attempt to order the elements of these sets. The systematization principle of the *E. octocarinatus* compressed code version is shown in Fig. 12a (Shcherbak, 1993a). According to the principle, an isolated triplet, an amino acid or a syntactic sign, is not an object of systematization. Two composite objects play this role. These objects are:

(i) the sets of the same degeneracy series;
(ii) the same degeneracy series within one of the sets.

Each of these objects is a component part of another one (see crossing arrows in Fig. 12a). The objects are aligned with a monotonic and parallel–opposition change of two parameters. These parameters are:

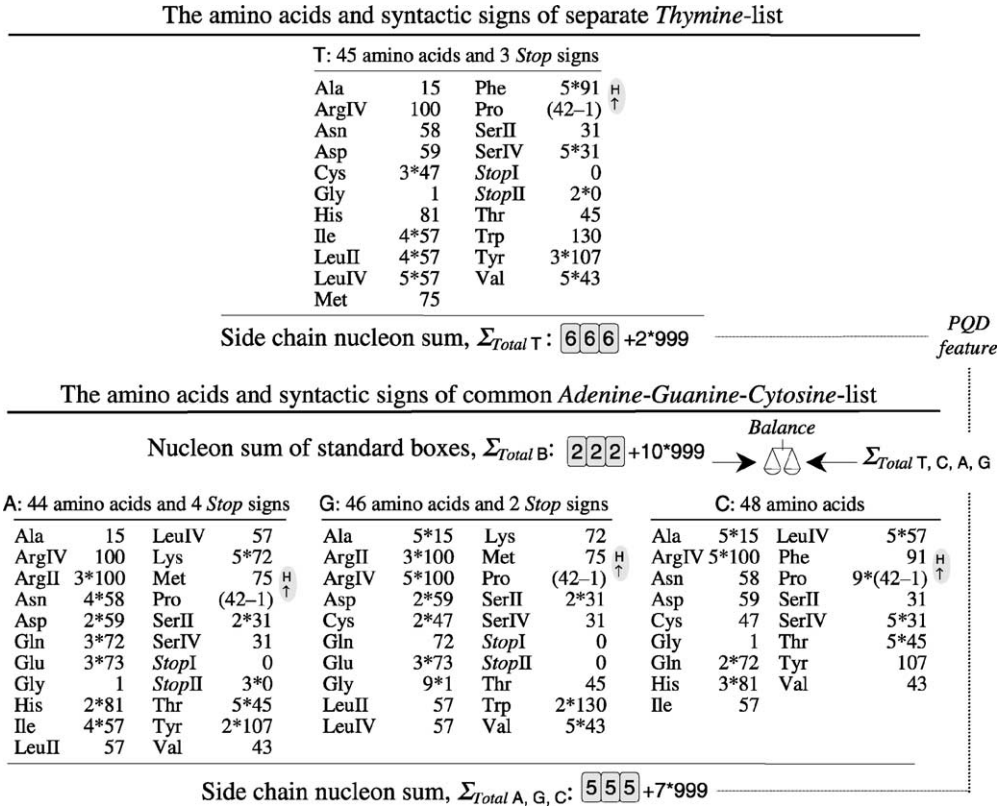


Fig. 11. The arithmetic regularity of the split representation of the genetic code (see Fig. 1c). The *Thymine*-list and the common *Adenine-Guanine-Cytosine*-list demonstrate the PQD features. These features indicate the block-and-chain balance between the amino acid side chains of the whole split genetic code and the fixed set of the standard blocks of the *A-G-C*-list. The revealed regularities are true for the *E. coli* version of the genetic code.

(i) the degeneracy number of the series triplets;
(ii) the nucleon number of the series amino acids (syntactic signs).

Essentially, the principle combines Rumer's division by the series degeneracy and Hasegawa and Miyata's series alignment by the masses into an organic whole (see Fig. 3 and Table 1). A reliable systematization principle must produce a strong and various ordering using a weak and uniform condition. Seemingly, the genetic code exceeds the fragmentary ASCII arrangement in this respect.

The word *calligramme* means spatially ordered symbols of the code elements. The calligramme in Fig. 12b shows the genetic code cooperative symmetry (Shcherbak, 1988, 1989a,b,c). The common parameters indicate that the arithmetic regularity and the cooperative symmetry are tightly bound phenomena.

It seems that a zero—fascinating invention of arithmetic—is in use by biocomputers. The zeros of the *Stop* signs were trivial summands in the nucleon sums of the arithmetic regularity. Now, the zero forms the calligramme as an *ordinal* number. On that ground, the zero occupies its formally predetermined position at the beginning of the natural number series and places its own triplet series at the flank of the degeneracy II set. Being in this position, these series entirely maintain cooperative symmetry. It is surprising that the zero, i.e. abstract nothing, operates the substantial objects of the code. However, a presumable implication of biocomputers to the modification of the genetic code offers a satisfactory explanation of that fact.

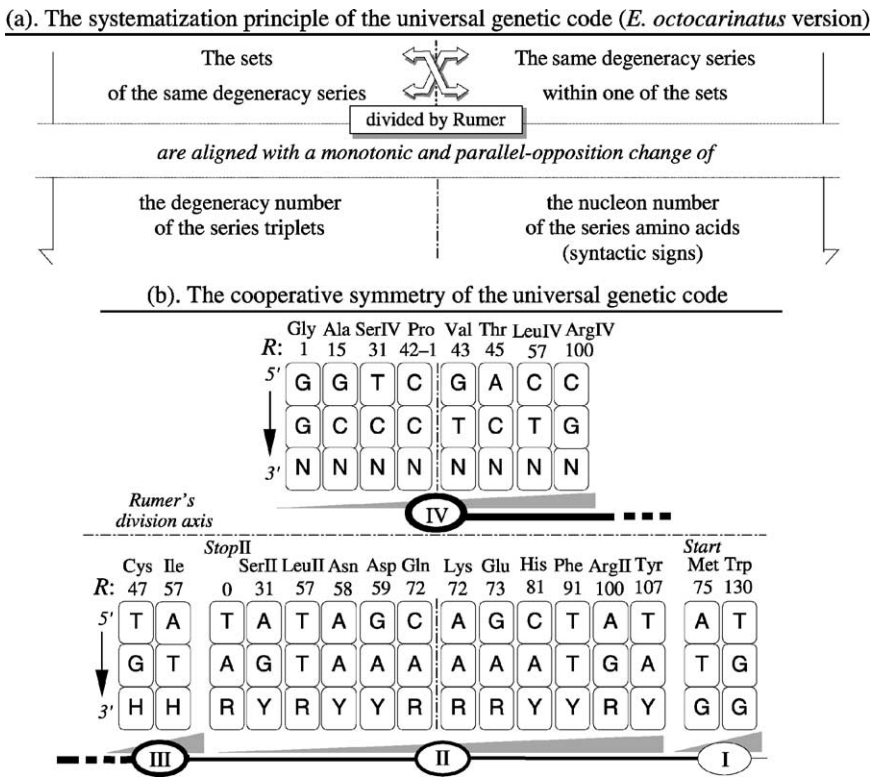


Fig. 12. The systematization principle of the genetic code. (a) The centrosymmetrical flow chart indicates equilibrium of the principle formulation. (b) The application of the principle to the compressed series in Fig. 1a results in the cooperative symmetry of the genetic code. The *calligramme* of the cooperative symmetry comprises entirely the regularity of the Rumer's division in Fig. 3 and the manifestation of the arithmetic regularity in Fig. 6. The lines of variable thickness and wedge-shaped figures symbolize an alignment with a monotonic and parallel-opposition change of the degeneracy and nucleon numbers.

9. The symmetry of the 5′ sequences

The degeneracy IV sequence is arranged by the mirror, translation, and inversion symmetry (Fig. 13a). Specifically, 5′ bases are invariant under a combined mirror and inversion operation; a pair of *RRYY* quartets forms a minimum pattern of the translation symmetry.

Three symmetry operations arrange the degeneracy III, II, I sequence, and in doing so the operations are the same as in the degeneracy IV set (Fig. 13b). A pair of flanking *TATAT* boxes is mirror symmetrical. A pair of the central *AGC* triplets forms a minimum pattern of the translation symmetry. The inversion and the absence of it, i.e. an identical transformation, interconnect the 5′ and 3′ bases of the degeneracy II set (Fig. 13c). This regularity is projected to the 3′ and 5′ bases of the tRNA antitriplets (Fig. 13d) and formalizes the *wobble pairing* rule (Crick, 1966). Since the tRNA antitriplets of the degeneracy II series have guanine as the purine 5′ base (Jukes, 1983), the order with regard to it is unambiguous. If this symmetry is maintained, there is no dilemma as to the relative position of amino acids Gln and Lys.

10. The symmetry of the *Center* sequences

Both sequences of the degeneracy IV (*RY*-representation) set and III, II, I set are mirror symmetrical (Fig. 14a). The sequence of the degeneracy IV set and its projection into the coaxially positioned sequence of the degeneracy III, II, I set are interconnected by the inversion symmetry. The coaxial symmetry is manifested due to the unique geometric shape of

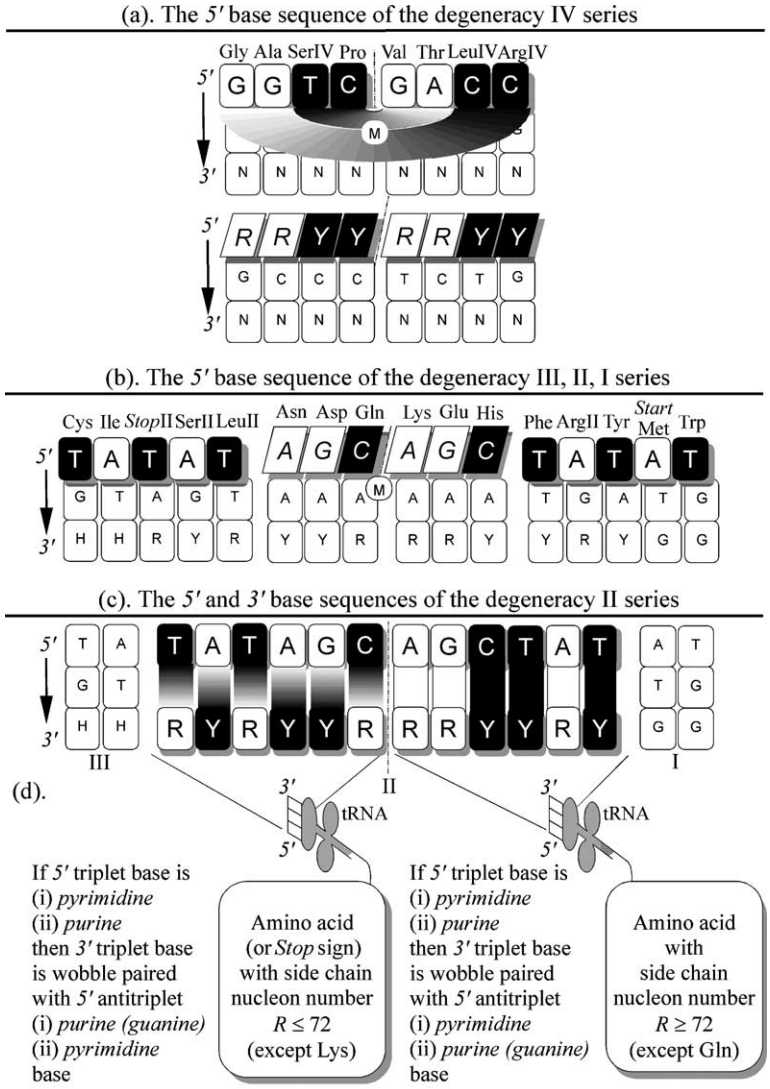


Fig. 13. The calligramme of the cooperative symmetry of the genetic code. Both 5′ base sequences are ordered by the common set of symmetry operations (a), (b), and (c). These are the mirror (designated by letter *M*), translation (designated by italicized letters *RY* or *TCAG*) and inversion of the *Base* → *Complementary base* or *Purine base* ↔ *Pyrimidine base* type (designated by color gradient whose white and black colors symbolize the purine and pyrimidine bases, respectively). (d) The *Purine base* ↔ *Pyrimidine base* inversion and identical transformation TCAG → TCAG formalize the wobble pairing of the 5′ and 3′ bases of the degeneracy II set. The graphic representation of the genetic code images its *E. octocarinatus* version; the revealed regularities are true for this version.

the calligramme. The shape and its coaxial symmetry decay in the scale of the life-size representation.

The classic mirror symmetry of the degeneracy IV set is replaced by its semantic variant for a palindromic feature (denoted by π) of two separate triplets in Fig. 14b. (The term *palindrome* is used in the linguistic sense as a line of characters, which are centrosymmetrical and whose semantics is invariant with respect to the directions of reading.) The symmetry-preserving palindromic feature is an indicator of similar semantic structure within the degeneracy III, II, I set. The point is that both *Center*

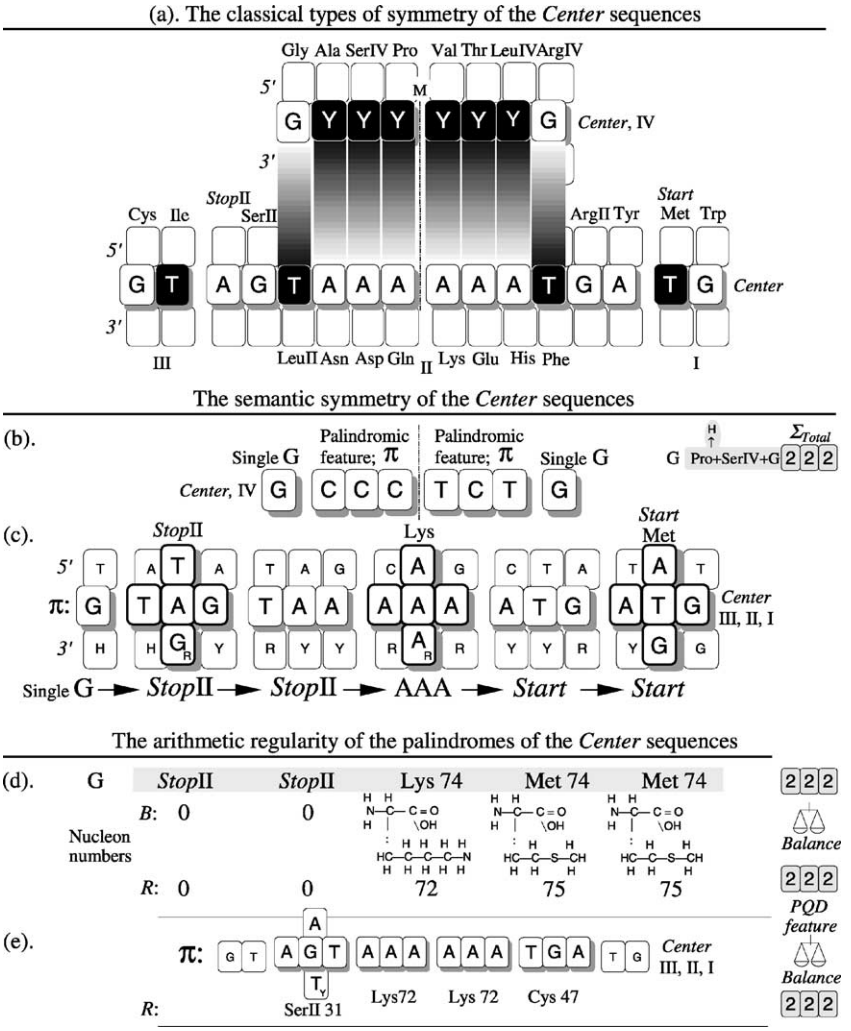


Fig. 14. The calligramme of the cooperative symmetry of the genetic code. Both sequences of *Center* bases are ordered by the common set of symmetry operations. These are the mirror (designated by *M*), the inversion of the *Purine base* ↔ *Pyrimidine base* type, and the semantic symmetry of the palindromic feature (designated by π). The graphic representation of the genetic code images its *E. octocarinatus* version; the revealed regularities are true for this version. The graphical symbols are the same as in Fig. 13.

sequences have, similar to the 5′ sequences above, a common set of symmetry operations, and the palindromic one is among them.

As predicted, the palindrome of the degeneracy III, II, I set is manifested in the triplet reading frame starting with the single base *G* at the flank (Fig. 14c). The palindrome has remarkable semantics expressed by the syntactic signs *Stop*II and *Start*. These signs are in a semantically antisymmetrical position with respect to the central homogeneous triplet *AAA*. The palindrome uses all syntactic signs of the *E. octocarinatus* code version: two termination signs *Stop*II and one twice-repeated initiation sign *Start*.

The unique palindrome semantics is not the only reason why the given position of the reading frame gets the better of its two other possible positions. This position is fixed as preferable also by a symmetric intersection between the real triplets and the abstract symbols of these triplets in the palindrome (shown in Fig. 14c by wide bordering). One of two directions

of reading is associated with the intersections and is preferable in this respect. The semantically arranged palindrome text, which is made by the code symbols within the code itself, is striking.

The arithmetic regularity appears surprisingly again, now through the palindrome semantics. The summation of nucleons reveals the block-and-chain balance and manifests the PQD 037 feature in Fig. 14b, d and e (the G residue in Fig. 14b has 150 nucleons).

The calligramme is adapted exactly to the *E. octocarinatus* code version. The series TGA-*Stop*I, absent in this version, is excluded from the syntactic signs of the palindrome, while its TGA-Cys substitute supports the manifestation of the PQD 037 feature in Fig. 14e.

The ASCII code comprises the arranged Roman letters and Arabic numeral series for the sake of the control of information. Similarly, the frequently used consensus sequences had the best chance to be incorporated into the calligramme. One can try to discern the symbols of the canonical base pairing in Fig. 13a, the TATA-box initiating transcription in Fig. 13b, a spacer zone with a poly(A) background between the *Stop* and the *Start* signs of neighboring genes in Fig. 14c, etc. One can assume from the symmetry consideration that the split *E. coli* code version in Fig. 11 keeps an unknown calligramme of $222 + 10 \times 999$ nucleons and 192 bases. The Pythagorean numbers (Fig. 6a), sequence *AGC* (Fig. 13b), the preferable directions (Fig. 14c), etc. could be its parameters.

11. The physicochemical effect of the cooperative symmetry

It is well known that the amino acids, triplet series, and degeneracy of the series are optimally distributed and, as a rule, possible consequences of the mutations such as an abrupt change of hydrophobicity, are, to some extent, weakened (Goldberg and Wittes, 1966; Volkenstein, 1966; Sjöström and Wold, 1985; Figureau, 1987; Knight et al., 1999). However, Shulz and Schirmer (1979) ask a question: why is it that another property of amino acids—a geometrical size of side chain—has no comparable protection though this property and is no less important than hydrophobicity?

One may conclude that the possession of cooperative symmetry does not allow the code to enhance the protection against an abrupt change of size. A calligramme fragment underlies the histograms of hydrophobicity and molecular size in Fig. 15. The transversions of the most mutable 3′ base are shown for the largest degeneracy II code set by arrows. Due to the symmetry of the fragment, the transversions are centrosymmetrical in their arrangement too. Recall that the triplet series are aligned by the monotonous change of the nucleon numbers of their amino acids. The molecular size of an amino acid is generally proportionate to its nucleon number. Therefore, the monotony excludes a centrosymmetrical shape of the size histogram. And vice versa, the hydrophobicity of an amino acid does not depend explicitly on its size and nucleon number. This allows the hydrophobicity to give any centrosymmetrical shape to the histogram, for example, such as that, which was fixed within the actual code. As a result, when amino acids are replaced through mutation the change of hydrophobicity is generally weak, while that of size is strong.

Freeland and Hurst (1998) showed that only one in every million random alternative codes generated is more efficient against mutation than the natural code. However, those codes hardly ever can get some comparable exact order at random. The natural genetic code is a unique one rather than one in a million in this respect. Thus, the natural code is highly efficient at minimizing the effects of mutation, though that ability is to some extent limited by arithmetic inside. It seems that the universal genetic code is arranged optimally.

12. The mathematical model of the genetic code

Ordering completeness and uniqueness needs a simple tool for evaluation. Diophantine equations and some conditions allow arithmetic regularity and cooperative symmetry to be described formally (Table 2). Such a system must show to what extent the code obeys formal ordering. In other words, it has to answer the question: could another arrangement of the canonical amino acids or other amino acid molecules having different nucleon numbers produce the same formal ordering?

The number of unknown variables is larger than that of the equations. This is typical of Diophantine

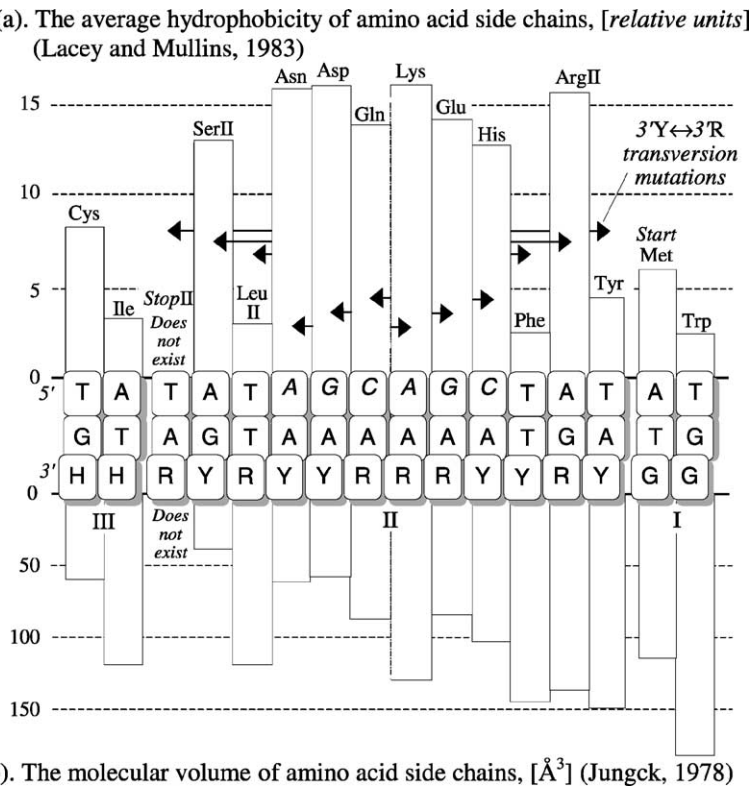


Fig. 15. Physicochemical effect of the cooperative symmetry of the genetic code is that if mutation occurs, the genetic code is satisfactorily protected against an abrupt change of hydrophobicity rather than the geometric size. The sequences of the 5′ and *Center* bases of the degeneracy III, II, I sets (see Figs. 13 and 14) are combined with the histograms of (a) hydrophobicity (Lacey and Mullins, 1983) and (b) molecular volume (Jungck, 1978). Arrows designate the transversion mutations of the 3′ position of the triplets as well as the corresponding substitutions of the amino acid hydrophobicity and geometric size within the degeneracy II set. This effect is true for both *E. coli* and *E. octocarinatus* code version. The graphic representation of the genetic code images its *E. octocarinatus* version.

equation systems, which have a solution owing to some side conditions (Table 3).

The roots of an algebraic system may be unique or contained within certain intervals depending to what extent the system is determined. The degree of determination becomes a reliable criterion of the evaluation of the code obeying the formal order.

The system was subjected to computational investigation and found completely defined. Its unique roots coincided with the nucleon numbers of the canonical amino acids and syntactic signs as well as with the actual degeneracy numbers. The $f{:}T \rightarrow A,S$ function maps uniquely these roots onto the life-size triplets of the analytical version of the genetic code (Table 4).

The roots are more than mere mathematics. These are 18 different nonzero numbers. They determine the necessary minimum of different amino acids so as yet the arithmetic regularity and cooperative symmetry could be realized.

The variable $B = 74$ limits the standard block to the configuration of the α-, but not β-amino acid, which is a formal solution of one of the problems of the code origin. It is obvious that a certain nucleon number is not always a symbol of a single molecule out of many natural amino acids. However, the root of the variable *Gly* equal to 1 does not correspond to any side chain, but that of the amino acid glycine. The same is true, at least, for the variables Ala, SerII, and SerIV. Thus, some roots are

Table 2
The Diophantine equation system comprises Eqs. (1)–(29) and logical conditions; there are auxiliary Eqs. (a) and (b) in this list

The arithmetic regularity of the compressed code representation in Fig. 6

Series compressed in scale $1:D_{\mathrm{IV}}$; $Gly + Ala + SerIV + Pro + Val + Thr + LeuIV + ArgIV = 333$; (a)

Series compressed in scales $1:D_{\mathrm{III}}$, $1:D_{\mathrm{II}}$, and $1:D_{\mathrm{I}}$; $(Ile) + (Cys + Stop\mathrm{II} + SerII + LeuII + Asn + Asp + G\mathrm{ln} + Lys + His + Phe + ArgII + Tyr) + (Stop\mathrm{I} + Met + Trp) = 111 + 999$. (b)

The arithmetic regularity of the compressed code representation in Figs. 6–8; the series with *Center* base either *C* or *A*

Compressed scale $1:D_{\mathrm{IV}}$; $5'T$; $SerIV = 31$; (1)

Compressed scale $1:D_{\mathrm{II}}$; $5'T$; $Stop\mathrm{II} + Tyr = 107$; (2)

Compressed scale $1:D_{\mathrm{IV}}$; $5'G$; $Ala = 15$; (3)

Compressed scale $1:D_{\mathrm{II}}$; $5'G$; $Asp + Glu = 132$; (4)

Compressed scale $1:D_{\mathrm{IV}}$; $5'C$; $Pro = 41$; (5)

Compressed scale $1:D_{\mathrm{II}}$; $5'C$; $Gln + His = 153$; (6)

Compressed scale $1:D_{\mathrm{IV}}$; $5'A$; $Thr = 45$; (7)

Compressed scale $1:D_{\mathrm{II}}$; $5'A$; $Asn + Lys = 130$; (8)

The arithmetic regularity of the compressed code representation in Figs. 6–8; the series with *Center* base either *T* or *G*

Compressed scale $1:D_{\mathrm{IV}}$; $5'C$; $LeuIV + ArgIV = 157$; (9)

Compressed scale $1:D_{\mathrm{IV}}$; $5'G$; $Gly + Val = 44$; (10)

Compressed scales $1:D_{\mathrm{III}}$; $1:D_{\mathrm{II}}$; $1:D_{\mathrm{I}}$; $5'T$; $LeuII + Phe + Cys + Stop\mathrm{I} + Trp = 325$; (11)

Compressed scales $1:D_{\mathrm{III}}$; $1:D_{\mathrm{II}}$; $1:D_{\mathrm{I}}$; $5'A$; $Ile + Met + SerII + ArgII = 263$. (12)

The scale symmetry of the compressed and life-size code representation

$(Gly + Ala + SerIV + Pro + Val + Thr + LeuIV + ArgIV) \times D_{\mathrm{IV}} + (Ile) \times D_{\mathrm{III}} + (Cys + Stop\mathrm{II} + SerII + LeuII + Asn + Asp + Gln + Lys + Glu + His + Phe + ArgII + Tyr) \times D_{\mathrm{II}} + (Stop\mathrm{I} + Met + Trp) \times D_{\mathrm{I}} = 3404$, i.e. $92PQ$. (13)

The arithmetic regularity of the life-size code representation in Figs. 9 and 10

Identical bases $(i) = \{Y\}$; unique bases $(u) = \{R\}$; $2LeuII + Ile + Tyr + Val + Cys + 2Pro + Thr + His + Ala + ArgIV = 691$; (14)

$i = \{Y\}$; $u = \{Y\}$; $Phe + 2LeuIV + 2SerIV + Pro = 308$; (15)

$i = \{A\}$; $5'$ or $3'u = \{Y\}$; $2Asn + Stop\mathrm{II} + Gln = 188$; (16)

$i = \{R\}$; $Center\ u = \{Y\}$; $i = \{N\}$; $5'u = \{R\}$; $Ile + Thr + Val + Ala = 160$; (17)

$i = \{G\}$; $5'$ or $3'u = \{Y\}$; $2Gly + Trp + ArgIV = 232$; (18)

$i = \{A\}$; $5'$ or $3'u = \{R\}$; $Lys + Glu = 145$; (19)

$i = \{G\}$; $5'$ or $3'u = \{R\}$; $Gly + ArgII = 101$; (20)

$i = \{R\}$; $Center\ u = \{R\}$; $i = \{N\}$; $5'u = \{R\}$; $ArgII + Glu = 173$. (21)

Triplets with either three identical or three unique bases; $Phe + Lys + SerIV + Ile + His + 2Ala + LeuIV + Cys + Stop\mathrm{I} + Asp + Met + Gln + SerII = 703$. (22)

The arithmetic regularity of the split code representation in Fig. 11

Triplets with base(s) *T*; $Ala + ArgIV + Asn + Asp + 3Cys + Gly + His + 4Ile + 4LeuII + 5LeuIV + Met + 5Phe + Pro + SerII + 5SerIV + Stop\mathrm{I} + 2Stop\mathrm{II} + Thr + Trp + 3Tyr + 5Val = 666 + 2 \times 999$. (23)

The arithmetic regularity of the palindromic features in Fig. 14d and e

$Stop\mathrm{II} + Stop\mathrm{II} + Lys + Start + Start = 222$; (24)

$SerII + 2Lys + Cys = 222$. (25)

The standard block nucleon sums of the block-and-chain balances in Figs. 6b, 7a, 11, and 14d

Series compressed in scales $1:D_{\mathrm{III}}$, $1:D_{\mathrm{II}}$, $1:D_{\mathrm{I}}$; $15B + b_{Stop\mathrm{II}} + b_{Stop\mathrm{I}} = 111 + 999$; (26)

Compressed series with $5'Y$; $10B + B_{Pro} + b_{Stop\mathrm{II}} + b_{Stop\mathrm{I}} = 814$; (27)

A-G-C-list of separate bases; $127B + 11B_{Pro} + 4b_{Stop\mathrm{II}} + 2b_{Stop\mathrm{I}} = 222 + 10 \times 999$; (28)

Compressed series of palindrom; $b_{Stop\mathrm{II}} + b_{Stop\mathrm{II}} + B + b_{Start} + b_{Start} = 222$. (29)

The system includes 33 positive integer variables. The majority of them correspond to the nucleon numbers of the amino acid side chains and standard blocks. These are the 23 chains designated as italicized three-letter abbreviations; two chains of the syntactic signs *Start* and $Stop\mathrm{I} = Stop\mathrm{II}$ (the same functionality equates both *Stop* variables); two blocks of the amino acid *B* and imino acid B_{Pro}; and two blocks of the syntactic signs b_{Start} and $b_{Stop\mathrm{I}} = b_{Stop\mathrm{II}}$. There are also four variables symbolizing the degeneracy numbers D_{IV}, D_{III}, D_{II}, D_{I} in this list. The logical conditions of the code division are a part of the model conditions. They are written in smaller fonts at the defined equations.

Algebra allows one to specify a minimum set of irreplaceable fragments necessary for composing from them the genetic code with arithmetic inside. The axis of Fig. 2a divides the compressed code into two fragments that are described by auxiliary Eqs. (a), (b), and (26). The simultaneous division by this and the other three axes of Fig. 2 produces a set of fragments described by Eqs. (1)–(12), (26), and (27).

The compressed and the life-size representations of the code are combined by the scale symmetry (Fig. 1). An algebraic symbol of that is Eq. (13). The life-size triplets with unique (*u*) and identical (*i*) bases demonstrate the arithmetic regularity in Fig. 9. The combined axes of Fig. 9 divide the code into eight fragments described by Eqs. (14)–(21). The regularities in Figs. 10 and 11 are described by Eqs. (22), (23), and (28).

The arithmetic regularity of the palindrome (Fig. 14) is described by Eqs. (24), (25), and (29).

The unambiguous assignment of the nucleon numbers of four various blocks through four Eqs. (26)–(29) is obvious. But it is not obvious a priori that there are enough balances of the block-and-chain types in the genetic code so that that assignment takes place.

Table 3
Side conditions of the Diophantine equation system

Integer numbers of degeneracy; $D_{IV} > D_{III} > D_{II} > D_{I} > 0$.	(30)
Series compressed in scale 1 : D_{IV}; $0 \leq Gly < Ala < SerIV < Pro < Val < Thr < LeuIV < ArgIV$.	(31)
E. octocarinatus series compressed in scale 1 : D_{III}; $0 \leq Cys < Ile$.	(32)
E. octocarinatus series compressed in scale 1 : D_{II}; $0 \leq Stop\text{II} < SerII < LeuII < Asn < Asp < Gln < Glu$;	(33a)
$Asp < Lys < Glu < His < Phe < ArgII < Tyr$.	(33b)
E. octocarinatus series compressed in scale 1 : D_{I}; $0 \leq Start < Trp$.	(34)
$\rceil \exists x \in [2, 14] \cup [16, 23] \cup [32, 38]$.	(35)
$StopI = StopII$	(36)

The sequences of inequalities (30)–(34) are the algebraic symbol of the cooperative symmetry. The classic and semantic symmetries are mapped onto these sequences in Figs. 13 and 14. These and other system conditions determine the mapping function $f{:}T \rightarrow A,S$ of the analytical version of the genetic code. For example, the palindrome semantic antisymmetry and symmetric intersections in Fig. 14c map both syntactic function *Stop*II and *Start* onto particular triplets and roots x_i.

Lehman and Jukes (1988) assumed that originally some triplets could temporarily lose encoded amino acids and turn into termination signs. The cooperative symmetry should extrude such triplets to the flanks of the four sets in Fig. 12b. For this reason, the inequalities (30)–(34) allow the flank variables *Gly*, *Cys*, *Start*, and *Met*, to have a relic zero, and vice versa, the variables *Stop*I = *Stop*II, to have non-zero nucleon numbers.

The canonical molecules together with other natural amino acids are incapable of forming an uninterrupted natural number series by nucleon numbers. Gaps, which are predicted reliably at least for small amino acids (Greenstein and Winitz, 1991), are assigned by condition (35).

associated with certain molecules numerically assigned to them.

The possession of the syntax signs *Stop*I and *Stop*II in the form of the standard triplet series is fraught with the danger of early terminations. Lewin (1987) assumed that for some reason the genetic code could not push two dangerous triplet series outside of its bounds in order to replace them with some non-triplet

Table 4
The analytical version of the genetic code shows the system solution in the form of the standard table

5′ position	*Center* position								3′ position
	T		*C*		*A*		*G*		
T	*TTT*	091	*TCT*	031 (Ser)	*TAT*	107	*TGT*	047	*T*
	TTC	091	*TCC*	031 (Ser)	*TAC*	107	*TGC*	047	*C*
	TTA	057	*TCA*	031 (Ser)	*TAA*	000 (*Stop*)	*TGA*	000[a] (*Stop*); 047[b]	*A*
	TTG	057	*TCG*	031 (Ser)	*TAG*	000 (*Stop*)	*TGG*	130	*G*
C	*CTT*	057	*CCT*	042–1	*CAT*	081	*CGT*	100	*T*
	CTC	057	*CCC*	042–1	*CAC*	081	*CGC*	100	*C*
	CTA	057	*CCA*	042–1	*CAA*	072	*CGA*	100	*A*
	CTG	057	*CCG*	042–1	*CAG*	072	*CGG*	100	*G*
A	*ATT*	057	*ACT*	045	*AAT*	058	*AGT*	031 (Ser)	*T*
	ATC	057	*ACC*	045	*AAC*	058	*AGC*	031 (Ser)	*C*
	ATA	057	*ACA*	045	*AAA*	072	*AGA*	100	*A*
	ATG	075 (*Start*)	*ACG*	045	*AAG*	072	*AGG*	100	*G*
G	*GTT*	043	*GCT*	015 (Ala)	*GAT*	059	*GGT*	001 (Gly)	*T*
	GTC	043	*GCC*	015 (Ala)	*GAC*	059	*GGC*	001 (Gly)	*C*
	GTA	043	*GCA*	015 (Ala)	*GAA*	073	*GGA*	001 (Gly)	*A*
	GTG	043	*GCG*	015 (Ala)	*GAG*	073	*GGG*	001 (Gly)	*G*

The parenthetical amino acid or syntactic sign imply that the numerical value of the corresponding root predetermines the chemical structure of amino acid or its absence in case of the termination. The PQD feature of the nucleon sum at the levels C_{β}, C_{δ}, etc., and the geometrically equivalent levels of the ring amino acids in Figs. 6 and 7a allows one to increase in number the parenthetical amino acids.

[a] The numerical symbol of the *E. coli* code version.

[b] The numerical symbol of the *E. octocarinatus* code version.

consensus sequences there. It seems that the formally indispensable zero roots of the system could be such a reason.

The nonzero root of the variable *Start* and zero roots of the variables *Stop*I and *Stop*II formally predetermine the essential differences of the initiation and termination, namely, the participation or absence of an amino acid in these acts. This may be an allusion to why the current translation has enough standard intermediates for the initiation instead of the specific factors for the termination.

Both regular degeneracy and universal triplet are indispensable conditions of the model. One can speculate that these phenomena owe their origin to this circumstance.

13. Discussion

Let us assume that at early stages of the origin of life it developed the arithmetic power of evaluative genetic language. The material objects of that power could be some molecular organelles working like biocomputers. These organelles should modify previous versions of the primordial genetic code in accord with a selected numeration system and some computing algorithms, for example, such as serial additions (Mac Dynaill, 2002). Then, the actual genetic code is a final variant of that modification.

Which of the existing models best describes the above events? Arithmetic provides specific requirements for selection. First, the arithmetic symbols need a freedom of the mapping function to appear into the code (Yockey, 2000). Second, any given collective regularity should be a jump phenomenon rather than a result of some gradual changes. Third, a suitable model should give an opportunity to both arithmetic regularity and physicochemical protection to arise.

Crick et al. (1976) introduced a primordial tRNAs to fix the product of the "frozen accident" (Crick, 1968). These intermediaries give desired freedom to the mapping function. They release hydrophobicity from hypothetical task to be direct and selective connector of triplets and amino acids. Porschke's experimental data (1985) did not show adequate recognition, accuracy, and attractive power to consider hydrophobicity as a unique and reliable code maker. The "released" hydrophobicity could take the protective shape in Fig. 15a with less difficulty. In addition, the tRNA intermediaries promote discontinuous jump of the primordial genetic code.

The code might have influenced and been influenced by the grammatical categories of the evaluative genetic language and arithmetic might be among these categories. The genetic machinery relates to discrete movement of different molecules in space, and consequently, an inertia effect of mass. One can suppose that this effect selected the massive nucleons as the amino acid's digital signs (Kashkarov et al., 2002).

Gamow (1954) speculated that hereditary properties could be characterized by a sequence of quaternary digits. That sequence represented by three-digit numbers could have at certain distances some checksums divisible e.g. by quaternary PQ 013_4. Theoretically, such arithmetic background may underlie any given genetic text without limitation of its biological content. Analyzing its own arithmetic, the background might check, restore, and alter superimposed biological content by means of calculations omitting the translation. One can speculate that such total analysis takes place in conjugate chromosomes during enigmatic meiosis prophase I, whose duration is up to 3 weeks for human male beings.

The divisibility by PQ as a validation criterion, if any, simplifies molecular machinery and facilitates the computational procedure of hypothetical organelles working as biocomputers. Perhaps a biocomputer is hardly more complicated than, e.g. mRNA-polymerase, and probably some known organelles have the capability to perform arithmetic.

A cell should get a computing result as a sequence of digits on some register. One may expect that the molecular register of biocomputers should be compact and highly accurate. Indeed, the test of divisibility by PQ of any given number independently of its value requires only the three-digit register. If the result is correct, the PQD feature gives its symmetrical shape to the molecular digits on the register. A routine shape analysis could convert these digits to an "analogue" form, and initiate certain biochemical reactions depending on the computing result. The symmetric shape or its absence contributes to a unique identifiability of the correct or incorrect register status. Note that a simple algorithm allows one to organize arithmetic division by PQ by summation of digits of a dividend number (see Fig. 5).

Although the revealed ordering does not go beyond the boundary of the genetic code, one can consider it as a universal pattern applicable to actual genetic texts. On the other hand, one can assume that only divisibility by the number 37 might be a final effect whose PQD 037 feature is only a by-product, which is indifferent for the genetics, and discernible by the human mind whose system is also the decimal one. If that is the case, some things of the strict physical nature such as the balances and structures of the cooperative symmetry remain valid as yet.

Acknowledgements

Part of this study was made during the author's visit to Max-Planck-Institute für biophysikalische Chemie, Göttingen, Germany. The author is greatly indebted to Professor Manfred Eigen for his support at that early stage of the present work and expresses his special thanks to Ruthild Winkler-Oswatitsch.

The work was financed by the Ministry of Science and Education of Republic of Kazakhstan and Max-Planck-Gesellschaft, Germany.

References

Adleman, L.M., 1994. Molecular computation of solutions to combinatorial problems. Science 266, 1021–1023.

Bertman, M.O., Jungck, J.R., 1979. Group graph of the genetic code. J. Heredity 70, 379–384.

Crick, F.H.C., 1966. Codon-anticodon pairing: the wobble hypothesis. J. Mol. Biol. 19, 548–555.

Crick, F.H.C., 1968. The origin of the genetic code. J. Mol. Biol. 38, 367–379.

Crick, F.H.C., Brenner, S., Klug, A., Pieczenik, G., 1976. A speculation on the origin of protein synthesis. Origins Life 7, 389–397.

Chen, J., Wood, D.H., 2000. Computation with biomolecules. Proc. Natl. Acad. Sci. U.S.A. 97, 1328–1330.

Dankwerts, H.J., Neubert, D., 1975. Symmetries of the genetic code doublets. J. Mol. Evol. 5, 327–332.

Eigen, M., Winkler, R., 1985. Das Spiel. Piper Verlag, München, Zürich, pp. 281–316.

Faulhammer, D., Cukras, A.R., Lipton, R.J., Landweber, L.F., 2000. Molecular computation: RNA solutions to chess problems. Proc. Natl. Acad. Sci. U.S.A. 97, 1385–1389.

Figureau, A., 1987. Information theory and the genetic code. Origins Life 17, 439–449.

Freeland, S.J., Hurst, L.D., 1998. The genetic code is one in a million. J. Mol. Evol. 47, 238–248.

Gamow, G., 1954. Possible relation between deoxyribonucleic acid and protein structures. Nature 13, 318.

Goldberg, A.L., Wittes, R.E., 1966. Genetic code: aspects of organization. Science 153, 420–424.

Greenstein, J.P., Winitz, M., 1991. Chemistry of the Amino Acids. John Wiley and Sons, Inc., New York, London.

Grimm, M., Brünen-Nieweler, C., Junker, V., Heckmann, K., Beier, H., 1998. The hypotrichous ciliate *Euplotes octocarinatus* has only one type of $tRNA^{Cys}$ with GCA anticodon encoded on a single macromolecular DNA molecule. Nucleic Acids Res. 26, 4557–4565.

Hasegawa, M., Miyata, T., 1980. On the antisymmetry of the amino acid code table. Origins Life 10, 265–270.

Jiménez-Montaño, M.A., 1993. On the syntactic structure and redundancy distribution of the genetic code. BioSystems 32, 11–23.

Jiménez-Montaño, M.A., 1999. Protein evolution drives the evolution of the genetic code and *vice versa*. BioSystems 54, 47–64.

Jiménez-Montaño, M.A., de la Mora-Basañez, C.R., Poeschel, T., 1996. The hypercube structure of the genetic code explains conservative and non-conservative amino acid substitutions *in vivo* and *in vitro*. BioSystems 39, 379–384.

Jukes, T.H., 1983. Evolution of the amino acid code: inferences from mitochondrial codes. J. Mol. Evol. 19, 219–225.

Jukes, T.H., Osawa, S., 1990. The genetic code in mitochondria and chloroplasts. Experimentia 46, 1117–1126.

Jungck, J.R., 1978. The genetic code as periodic table. J. Mol. Evol. 11, 211–224.

Karasev, V.A., Sorokin, S.G., 1997. Topological structure of the genetic code. Russian J. Genetics 33, 622–628.

Kashkarov, V.V., Krassovitskiy, A.M., Mamleev, V.S., shCherbak, V.I., 2002. Random sequences of proteins are exactly balanced like the canonical base pairs of DNA. In: Proceedings of the 10th ISSOL Meeting and 13th International Conference on the Origin of Life, Oaxaca City, Mexico, June 30–July 4.

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, 241–247.

Konopelchenko, B.G., Rumer, Yu.B., 1975. The codon classification inside genetic code. Docl. Acad. Nauk. SSSR 223, 471–474.

Lacey Jr., J.C., Mullins Jr., D.W., 1983. Experimental studies related to the origin of the genetic code and the process of protein synthesis—a review. Origins Life 13, 3–42.

Landweber, L.F., Gilbert, W., 1993. RNA editing as a source of genetic variation. Nature 363, 179–182.

Landweber, L.F., Kuo, T.-C., Curtis, E., 2000. Evolution and assembly of an extremely scrambled gene. Proc. Natl. Acad. Sci. U.S.A. 97, 3298–3303.

Lehman, N., Jukes, T.H., 1988. The genetic code development by stop codon takeover. J. Theor. Biol. 135, 203–214.

Lewin, B., 1987. Genes. John Wiley and Sons, Inc., New York, London.

Mac Dynaill, D.A., 2002. A parity code interpretation of nucleotide alphabet composition. Chem. Commun. 18, 2062–2063.

Marshal, R.E., Cascey, T.C., Nirenberg, M., 1967. Fine structure of RNA codewords recognized by bacterial, amphibian and mammalian transfer RNA. Science 155, 820–825.

Meyer, F., Schmidt, H.I., Plümper, E., Hasilik, A., Mersmann, G., Meyer, H.E., Engström, A., Heckmann, K., 1991. UGA is translated as cysteine in pheromone 3 of *Euplotes octocarinatus*. Proc. Natl. Acad. Sci. U.S.A. 88, 3758–3761.

Porschke, D., 1985. Differential effect of amino acid residues on the stability of double helices formed from polyribonucleotides and its possible relation to the evolution of the genetic code. J. Mol. Evol. 21, 192–198.

Rosen, G., 1999. Molecular content relations in the genetic code. Phys. Lett. A 253, 354–357.

Rumer, Yu.B., 1966. About systematization of the codons of the genetic code. Dokl. Acad. Nauk. SSSR 167, 1393–1394.

Shcherbak, V.I., 1988. The co-operative symmetry of the genetic code. J. Theor. Biol. 132, 121–124.

Shcherbak, V.I., 1989a. Rumer's rule and transformation in the context of the co-operative symmetry of the genetic code. J. Theor. Biol. 139, 271–276.

Shcherbak, V.I., 1989b. Ways of wobble pairing are formalized with the co-operative symmetry of the genetic code. J. Theor. Biol. 139, 277–281.

Shcherbak, V.I., 1989c. The "START" and "STOP" of the genetic code: why exactly ATG and TAG, TAA? J. Theor. Biol. 139, 283–286.

Shcherbak, V.I., 1993a. The symmetrical architecture of the genetic code systematization principle. J. Theor. Biol. 162, 395–398.

Shcherbak, V.I., 1993b. Twenty canonical amino acids of the genetic code: the arithmetical regularities. Part I. J. Theor. Biol. 162, 399–401.

Shcherbak, V.I., 1994. Sixty-four triplets and 20 canonical amino acids of the genetic code: the arithmetical regularities. Part II. J. Theor. Biol. 166, 475–477.

Shcherbak, V.I., 1996. A new manifestation of the arithmetical regularity suggests the universal genetic code distinguishes the decimal system. In: Proceedings of the 11th International Conference Origin of Life, Book of Abstracts, Orleans July 5–12, France.

Shcherbak, V.I., 1999. A new manifestation of the decimal system in the genetic code. In: Proceedings of the 12th International Conference Origin of Life, Book of Abstracts, San-Diego, July 11–16, USA.

Shulz, G.E., Schirmer, R.H., 1979. Principles of Protein Structure. Springer-Verlag, New York, Heidelberg, Berlin.

Siemion, I.Z., 1994. Compositional frequencies of amino acids in the proteins and the genetic code. BioSystems 32, 163–170.

Sjöström, M., Wold, S., 1985. A multivariate study of the relationship between the genetic code and the physical–chemical properties of amino acids. J. Mol. Evol. 22, 272–277.

Sukhodolets, V.V., 1989. The genetic code as a clue of molecular evolution. J. Theor. Biol. 141, 379–389.

Tan, M., Heckmann, K., 1998. The two γ-tubulin-encoding genes of the ciliate *Euplotes crassus* differ in their sequences, codon usage, transcription initiation sites and poly(A) addition sites. Gene 210, 53–60.

Tan, M., Heckmann, K., Brünen-Nieweler, C., 1999. The micronuclear gene encoding a putative aminoacyl-tRNA synthetase cofactor of the ciliate *Euplotes octocarinatus* is interrupted by two sequences that are removed during macromolecular development. Gene 233, 131–140.

Verkhovod, A.B., 1994. Alphanumerical division of the universal genetic code: new divisions reveal new balances. J. Theor. Biol. 170, 327–330.

Volkenstein, M.V., 1966. The genetic code of protein structures. Biochim. Biophys. Acta 119, 421–424.

Yockey, H.P., 1992. Information Theory and Molecular Biology. Cambridge University Press, Cambridge, New York.

Yockey, H.P., 2000. Origin of life on earth and Shannon's theory of communication. Comput. Chem. 24, 105–123.

Zhang, C.T., 1997. A symmetrical theory of DNA sequences and its applications. J. Theor. Biol. 187, 297–306.