[GADV]-amino Acids and GNC Codons Selected

Subjects: Biochemistry & Molecular Biology

Contributor: Kenji Ikehara

The genetic code connecting gene with protein is one of the six members (gene, tRNA, genetic code, protein, metabolism, and cell structure) composing the fundamental life system, including the core life system (gene, tRNA, genetic code, and protein). Revealing the origin of the genetic code, one of four members of the core life system might lead to solving the mystery of the origin of life. Therefore, it is quite important to clarify how amino acids and codons were selected among messy organic compounds on primitive Earth and how the first genetic code was established.

Keywords: origin of the genetic code ; GNC code frozen-accident theory ; coevolution theory

1. GNC-SNS Primitive Genetic Code Hypothesis

1.1. GC-NSF(a) Hypothesis for Formation of Entirely New Genes

The study on the GADV hypothesis started from the origin of modern genes independently of the origins of the genetic code and life. Consequently, the GC-NSF(a) hypothesis was obtained, which suggests that entirely new genes are created from nonstop frames on GC-NSF(a)s. Note that those studies were carried out with amino acid composition and six properties of the respective amino acids, or hydropathy, α-helix, β-sheet, turn/coil formabilites, and acidic and basic amino acid compositions, all of which are related to water-soluble globular protein formation.

1.2. SNS Primitive Genetic Code Hypothesis

Next, base compositions at three codon positions were analyzed to reveal the reason why entirely new genes could be generated from an essentially random codon sequence encoded by GC-NSF(a). From the results, it was found that G and C are naturally used at high frequencies at the first and the third codon positions in the GC-rich region, but curiously, four bases, G, C, A, and T, are used at similar frequencies at the second base position of codon of GC-rich genes and that, in other words, codon sequences similar to $(SNS)_n$ are written on antisense strands of GC rich genes [1].

It was further analyzed, whether or not imaginary proteins encoded by SNS code, which were generated on a computer, can satisfy the six conditions (hydrophobicity/hydrophilicity or hydropathy, α-helix, β-sheet, turn/coil formabilities, and acidic and basic amino acid compositions) for water-soluble globular protein formation.

1.3. GNC Primeval Genetic Code Hypothesis

Furthermore, I explored whether or not more primitive imaginary proteins, which use only four amino acids extracted from columns and rows of the universal genetic code, can satisfy the four conditions (hydropathy, α-helix, β-sheet, and turn/coil formabilites), in which two other conditions (acidic and basic amino acid compositions) are excluded because of less importance than others. It became clear that essentially only one code composed of four GNC codons and four [GADV]-amino acids could satisfy the four conditions, except GNG code, in which Glu is used instead of Asp and encodes four [GAVE]-amino acids. Then, I presented the GNC-SNS primitive genetic code hypothesis [2].

2. Why Were [GADV]-amino Acids and GNC Codons Selected?

2.1. How Were Four [GADV]-amino Acids Selected?

It is well known that various amino acids, including simple β-amino acids and γ-amino acids, can be produced other than natural amino acids under experimental conditions carried out by Miller [3]. It is also confirmed by experiments carried out by Cleaves et al., in which oxidative gas was used instead of the reductive gas used in Miller's experiments, that various amino acids were similarly synthesized [4]. Therefore, it becomes a problem how [GADV]-amino acids could be selected out from the messy amino acids, which accumulated on primitive Earth at large amounts. I will give an answer to the problem below.

2.1.1. Why Are Only α-amino Acids Used in Proteins?

Only α-amino acids are used in the universal genetic code. The reason is because the use of only α-amino acids restricts free rotation to two bonds neighboring a carbon atom, which are sandwiched in between two peptide bonds having considerable double-bond character which prevents free rotation. On the contrary, use of β-amino acids and γ-amino acids make it difficult to form regular structures as α-helix and β-sheet structures because of additional free rotations of chemical bonds between α-carbon and β-carbon atoms and between β-carbon and γ-carbon atoms. Therefore, only α-amino acids must be used in protein synthesis. Inversely stating this, proto-cells using only α-amino acids could have a high productivity and left many descendants.

2.1.2. Why Are Twenty Natural Amino Acids Used in Proteins?

It is also well known that various types of simple α-amino acids as 2-aminobutylic acid (2-ABA: α-amino-n-butylate) and norvaline are also synthesized in Miller's experiments [3]. However, it has previously been unknown why those nonnatural α-amino acids are not used in proteins. The reason is explained below.

Amino acids having more than two methyl groups or a bulky side chain like a benzene ring of phenylalanine on β-carbon atom should have a large propensity for β-sheet formation because of a possible steric hindrance preventing α-helix formation. On the contrary, amino acids having two or three hydrogen atoms on a β-carbon atom should have a large propensity for α-helix formation. Therefore, use of 2-ABA or norvaline without any bulky side chain on a β-carbon atom instead of valine having two methyl groups on a β-carbon atom causes an unfavorable excess of α-helix formability and, in parallel, insufficient β-sheet formability of proteins synthesized under an imaginable GNC code encoding three [GAD]-amino acids+2-ABA or norvaline.

2.1.3. Why Are Only L-Amino Acids Used in Proteins?

Not only L-amino acids, but also D-amino acids, should be synthesized on primitive Earth without any asymmetrical field. Nevertheless, only L-amino acids are used in natural proteins. In other words, twenty amino acids are homochiral. In this case too, amino acids, which are used in proteins, must be homochiral. Secondary structures or regular structures are unsuccessfully formed if both L-amino acids and D-amino acids are used in protein synthesis, because propensity of secondary structure formation of D-amino acids is opposed to that of L-amino acids. On the other hand, Gly without asymmetric carbon atom is used for inhibiting secondary structure formation and promoting turn/coil formation of [GADV]-proteins.

The homochirality of three [ADV]-amino acids could be attained by differential crystallization, which accidentally occurred between racemic amino acid mixtures during the drying process of the [GADV]-amino acid aqueous solution under sunlight on primitive Earth [5][6].

2.1.4. Why Are Hydrophobic Val and Hydrophilic Asp Encoded in the GNC Code?

The existence of both hydrophobic Val, which forms the hydrophobic core structure in a protein, and hydrophilic Asp, which is favorable to locate on the surface of a protein, are indispensable to form a stable globular structure in water. Thus, usage of three L-α-[ADV]-amino acids plus Gly is necessary to form water-soluble globular proteins. Furthermore, it was confirmed that four [GADV]-amino acids are the simplest combination among twenty natural amino acids, with which water-soluble globular protein can be effectively formed [7]. Therefore, cell structures using four [GADV]-amino acids were selected as the results of repeated "trial and error" or of natural selection among the proteins using various types of amino acids, which accumulated on the primitive Earth in large amounts.

Naturally, proteins before formation of genetic information must be produced through random processes. Therefore, the most primitive proteins must be produced by the direct random joining of [GADV]-amino acids under one of protein 0th-order structures [8]. That is the only way for the production of immature but meaningful proteins leading to the emergence of life. This was confirmed by analysis, in which points were given when imaginary [GADV]-proteins satisfied the four conditions for water-soluble globular protein synthesis [2][7][9].

It is well known that [GADV]-amino acids are easily synthesized by Miller-type experiments [4][5] and are detected at large amounts from Murchison meteorite [10][11]. Therefore, the GADV hypothesis is consistent with the results obtained by Miller-type experiments and the results of chemical analyses of the meteorites, although the hypothesis is not founded on sufficient experimental results at this point in time. Further, it is also confirmed that [GADV]-amino acid composition is the simplest combination among twenty amino acids, which satisfies the four conditions for water-soluble globular protein formation [7].

Thus, steps to the emergence of life began upon the formation of immature [GADV]-proteins. Formations of GNC primeval genetic code and the first $(GNC)_n$ gene succeeded the immature [GADV]-protein formation as aiming at more efficiently producing [GADV]-proteins with higher functionality step by step. The first genetic code and gene became established as the results of repeated "trial and error" or through selection of [GADV]-microspheres, which could grow and proliferate faster than before through acquisition of immature but more and more efficient [GADV]-proteins [9].

The GNC primeval genetic code hypothesis, suggesting from which the universal genetic code originated from the GNC code, is consistent with the idea, which was described in Watson's textbook [12], and in the results obtained by Shepherd [13]. Therefore, I am convinced that the universal genetic code originated from the GNC code.

However, the GNC code described above is only the result that was obtained by investigation in a frame of the universal genetic code, which was triggered by a study on where and how entirely new genes are generated in extant organisms. Therefore, the reason is not evident why GNC codons were used in the first genetic code.

2.2. How Were Four GNC Codons Selected for the First Genetic Code?

Next, consider the reason how four GNC codons were selected and used in the first genetic code as suggested by the GNC-SNS primitive genetic code hypothesis [2].

2.2.1. Grounds Showing That GNC Codons Were Used in the First Genetic Code

Of course, [GADV]-proteins produced by the direct random joining of [GADV]-amino acids is essentially the same as polypeptide chains synthesized with random $(GNC)_n$ codon sequences, because both the proteins and polypeptides have a random [GADV]-amino acid sequence.

On the other hand, it is impossible to consider the formation process of the first genetic code connecting codons with amino acids independently of tRNAs, which actually mediate between codons and amino acids. It is then explained how the first tRNA was generated and what anticodons were used in the first tRNA. Naturally, the first tRNA must also be formed through random processes in the absence of genes. In this regard, I have proposed anticodon stem-loop (AntiC-SL) tRNA hypothesis on origin of tRNA [14], suggesting that modern tRNAs originated from [GADV]-AntiC-SL tRNAs, which were formed as the smallest but sufficiently stable hairpin loop RNAs composed of seventeen nucleotides through repeated random joining of nucleotides and degradation of oligonucleotides. It is obvious that the AntiC-SL tRNAs are sufficiently stable against RNase activity of immature [GADV]-proteins, because AntiC-SLs of modern *Escherichia coli* L-form tRNAs carrying a GNC anticodon that are not chemically modified except Asp-tRNA modified with a small methl group (/) and queosine (Q) [15].

2.2.2. Strong Binding of a Triplet, GNC, with the Complementary Triplet, GNC

Furthermore, it is also known that complementary triplet pairs, ${}^{5'}GNC^{3'}/{}^{3'}CNG^{5'}$, are more stable than any other complementary triplet pairs except Ser, as described in the paper published by Taghavi et al. (2017) [16], that *a series of codons of the form GNC (where N means 'anything') should have the special property of partitioning naturally at the codon boundary C to G when under tension*. Their observations are supported by the fact that a GNC anticodon carried by an AntiC-SL binds with the complementary codon in mRNA during translation.

3. How Were the Correspondence Relations between GNC Codons and [GADV]-amino Acids Established?

3.1. Direct Complex Formation between GNC Anticodons/Codons and [GADV]-amino Acids Is Impossible

Here, it is considered whether or not complexes can be formed between GNC anticodons and [GADV]-amino acids, and whether or not the complexes can be used for [GADV]-protein synthesis, if the complexes could be formed.

(1) Triplet GNC codons could not be directly bound with the respective [GADV]-amino acids as suggested by the stereochemical theory, which was proposed by Shimizu [17], because of the sizes of triplet GNC nucleotides. Even when the triplets were folded into a compact tertiary structure, they were too large to bind with small side chains as H-atom of Gly and methyl group of Ala. This is supported from the results obtained by Yarus [18] that stereospecific interaction between a triplet codon and its cognate amino acid with a small side chain as Gly, Ala, and so on could not be detected in complexes of RNA with proteins as ribosomes and riboswitches [18].

(2) Amino group and/or carboxyl group of [GADV]-amino acids could not expose outside from the complexes between anticodons and the corresponding amino acids, even if triplet nucleotides could be bound with [GADV]-amino acids, because those groups must be used for stable complex formation with triplet nucleotides. This means that the amino group and/or carboxyl group of [GADV]-amino acids cannot participate in [GADV]-protein synthesis.

(3) In addition to that, the binding mode of four complexes of GNC codons with [GADV]-amino acids must be the same, because otherwise it becomes impossible to form a peptide bond between two neighboring [GADV]-amino acids in the complexes. This also indicates that it would be impossible to synthesize immature [GADV]-proteins having various amino acid sequences by using the complexes of GNC codons/anticodons with [GADV]-amino acids.

(4) Furthermore, if the direct complexes between GNC codons and [GADV]-amino acids were used for [GADV]-protein synthesis, the direct stereochemical recognition system using complexes between GNC codons and [GADV]-amino acids for protein synthesis must, one day, be transferred to the indirect protein synthetic system using tRNA. However, it would be impossible to transfer the direct system to the indirect system with tRNA [9].

As explained so far, it is considered at this point in time that four [GADV]-amino acids and four GNC codons were selected for establishment of the GNC code. However, the reason why the correspondence relations between GNC codons and [GADV]-amino acids were determined is still entirely unknown. Inversely stating this, if it is considered from the standpoint of stereochemical theory, it must be shown that a historical trajectory from the direct joining to indirect joining of [GADV]-amino acids under the first genetic code must be explained.

Therefore, the above considerations clearly indicate that the GNC primeval genetic code could not be established as the stereochemical theory assumes. In order to overcome the difficulties, it is important to understand how the correspondence relations between four GNC codons and four [GADV]-amino acids were determined during formation of the most primitive but specific AntiC-SL tRNAs. Then, I would like to give an answer to the second problem of how the GNC primeval genetic code was established, which is described in the introduction.

3.2. GNC Code Frozen-Accident Theory on the Origin of the Genetic Code

Then, how was the first GNC code established? During the above considerations, the GNC code freeze-accident theory was conceived, suggesting that only the correspondence relations between GNC codons and [GADV]-amino acids were accidentally determined and frozen after the correspondence relations were established [9]. The establishment process of the primeval genetic code could be consistently explained by incorporating the GNC code freeze-accident theory [9]. I firmly believe now that the GNC primeval genetic code was established as assumed by the GNC code frozen-accident theory [9], because it is considered that there is no other way. However, a problem still remains unsolved of how corresponding relations between GUC-Val, GCC-Ala, GAC-Asp, and GGC-Gly were formed, because there exists 24 combinations (4! = 24) between GNC codons and [GADV]-amino acids, even under the assumption that use of two pairs between Gly-Ala and Asp-Val were inevitable. The assumption is based on formation of entirely new [GADV]-protein synthesis, which should be carried out under the GNC code and $(GNC)_n$ genes. Therefore, the last problem about the origin of the genetic code is how the four combinations, GUC-Val, GCC-Ala, GAC-Asp, and GGC-Gly, were selected and used in the GNC code, which evolved to the universal genetic code. I consider the problem as follows.

3.3. How Was the First GNC Code Established?

In order to completely solve the problem of how the first GNC code was established, it is necessary to clarify how the correspondence relations of four GNC codons and four [GADV]-amino acids were selected among 24 ways of combinations. Here, researchers must pay enough attention to the fact that selection of combinations is not always completely random, because any mature [GADV]-proteins having a rigid and compact structure must always be formed from an immature [GADV]-protein with a flexible and swollen structure. Therefore, it means that two amino acid pairs, Val-Asp and Ala-Gly, must use either one of two complementary pairs, GUC-GAC and GCC and GGC. The combinations are restricted in eight ways.

References

1. Ikehara, K.; Amada, F.; Yoshida, S.; Mikata, Y.; Tanaka, A. A possible origin of newly-born bacterial genes: Significance of GC-rich nonstop frame on antisense strand. Nucl. Acids Res. 1996, 24, 4249–4255.
2. Ikehara, K.; Omori, Y.; Arai, R.; Hirose, A. A novel theory on the origin of the genetic code: A GNC-SNS hypothesis. J. Mol. Evol. 2002, 54, 530–538.

3. Miller, S.L.; Orgel, L.E. The Origins of Life on the Earth; Prentice-Hall: Englewood Cliffs, NJ, USA, 1974.
4. Cleaves, H.J.; Chalmers, J.H.; Lazcano, A.; Miller, S.L.; Bada, J.L. A reassessment of prebiotic organic synthesis in neutral planetary atmosphere. Orig. Life Evol. Biosph. 2008, 38, 105–115.
5. Kojo, K. Origin of homochirality of amino acids in the biosphere. Symmetry 2010, 2, 1022–1032.
6. Breslow, R.; Levine, M.S. Amplification of enantiomeric concentrations under credible prebiotic conditions. Proc. Natl. Acad. Sci. USA 2006, 103, 12979–12980.
7. Ikehara, K. Origins of gene, genetic code, protein and life: Comprehensive view of life system from a GNC-SNS primitive genetic code hypothesis. J. Biosci. 2002, 27, 165–186.
8. Ikehara, K. Protein ordered sequences are formed by random joining of amino acids in protein 0th-order structure, followed by evolutionary process. Orig. Life Evol. Biosph. 2014, 44, 279–281.
9. Ikehara, K. Towards Revealing the Origin of Life—Presenting the GADV Hypothesis; Springer Nature, Gewerbestrasse: Cham, Switzerland, 2021.
10. Higgs, P.G. A four-column theory for the origin of the genetic code: Tracing the evolutionary pathways that gave rise to an optimized code. Biol. Direct. 2009, 24, 4–16.
11. Van der Gulik, P.; Massar, S.; Gilis, D.; Buhrman, H.; Rooman, M. The First peptides: The evolutionary transition between prebiotic amino acids and early proteins. J. Theor. Biol. 2009, 261, 531–539.
12. Watson, J.D.; Hopkins, N.H.; Roberts, J.W.; Steitz, J.A.; Weiner, A.M. Molecular Biology of the Gene, 4th ed.; The Benjamin/Cummings Publishing Company, Inc.: Menlo Park, CA, USA, 1987; p. 94025.
13. Shepherd, J.C.W. Fossil remnants of a primeval genetic code in all forms of life? Trends Biochem. Sci. 1984, 9, 8–10.
14. Ikehara, K. The origin of tRNA deduced from Pseudomonas aeruginosa 5′ anticodon-stem sequence: Anticodon stemloop hypothesis. Orig. Life Evol. Biosph. 2019, 49, 61–75.
15. Transfer RNA Database (Universitat Leipzig). Available online: http//trnadb.bioinf.uni-leipzig.de (accessed on 25 October 2021).
16. Taghavi, A.; van der Schoot, P.; Berryman, J.T. DNA partitions into triplets under tension in the presence of organic cations, with sequence evolutionary age predicting the stability of the triplet phase. Q. Rev. Biophys. 2017, 50, e15.
17. Shimizu, M. Molecular basis for the genetic code. J. Mol. Evol. 1982, 18, 297–303.
18. Yarus, M. The genetic code and RNA-amino acid affinities. Life 2017, 7, 13.