 doi:10.20944/preprints201712.0170.v1

Review

GADV-protein: An alternative to RNA in the origin of life

Kenji Ikehara[1, 2] *

[1] G&L Kyosei Institute, Koharu Bld. 202, Hokkeji 153-4, Nara, Nara 630-8001, Japan. ; ikehara@cc.nara-wu.ac.jp
[2] The International Institute for Advanced Studies of Japan, Kizugawadai 9-3, Kizugawa, Kyoto 619-0225, Japan.
* Correspondence: ikehara@cc.nara-wu.ac.jp; Tel.: +81-774-73-4478

Abstract: All life on Earth uses three integrated molecular systems in which genetic information contained in DNA base sequences is transmitted to ribosomes by RNA and a genetic code, then translated into the amino acid sequences of structural and catalytic proteins. Therefore, the most important point for understanding the origin of life is to determine how such systems could emerge from random processes on the early Earth. In this review, two alternatives are compared: the RNA world hypothesis and the [GADV]-protein world hypothesis. [GADV] refers to four amino acids, Gly [G], Ala [A], Asp [D] and Val [V] that are conserved in the amino acid sequences of many common proteins. Here I will argue that the origins of the three primary processes required for life to begin can be better explained by the GADV hypothesis than the RNA world hypothesis. The GADV hypothesis also incorporates a conversion process by which random polymers can evolve into proteins with ordered sequences.

Keywords: RNA world; [GADV]-protein world; GADV hypothesis; Origin of life; Protein 0^{th}-order structure; Origin of protein; Origin of genetic code; Origin of gene

1. Introduction

Two major hypotheses, the RNA world hypothesis [1] and [GADV]-protein world hypothesis (GADV hypothesis) [2, 3], have been proposed to account for how the first system composed of genes, a genetic code and proteins was established at the time that life began on the early Earth. (GADV are the single letter abbreviations referring to four amino acids, glycine, alanine, aspartic acid and valine that will be discussed later.) Other hypotheses related to Panspermia, hydrothermal vents [4-6], amyloids [7], coenzymes [8] and the tRNA core [9] have been also presented. However, these will not be considered here because they do not take into account the most important point, which is how the first genetic system was established (Figure 1A).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

RNA world hypothesis and [GADV]-protein world hypothesis

For many years, the well-known "chicken and egg relationship" between genes and proteins was a problem for origin of life research. Francis Crick, Leslie Orgel and Carl Woese speculated that a solution may involve RNA, and the discovery of ribozymes [10,11] provided a foundation for the idea. Gilbert [1] proposed the term RNA World to describe a time in which the first forms of life incorporated ribozymes having both genetic and catalytic functions. It is also considered that the first genetic system later evolved the capacity to transfer sequence information on the ribozyme to DNA and then to catalytic functions of proteins (Figure 1B). Szostak and his group developed an elegant procedure called SELEX: Systematic Evolution of Ligands by Exponential enrichment) to discover aptamers [12,13] which can bind to a variety of biochemical compounds such as nucleic acids and proteins [14]. Furthermore, Szostak et al., used SELEX to produce new ribozymes from a large pool of random sequences [15,16]. The generation of artificial ribozymes strongly supported the RNA world concept which is now described in biochemistry textbooks.

Before the discovery of double-stranded DNA [17] proteins were considered to be the primary constituent of cells and perhaps even as genetic materials. However, when the RNA world hypothesis was proposed, the consensus was that life could not emerge from proteins alone. The reason is based on two assumptions: (1) Proteins cannot be produced without genes because amino acid sequence diversity is extraordinary large -- 20^{100} = ~10^{130} -- even for a small protein composed of 100 amino acids [18]. (2) Proteins cannot be replicated because amino acids are unable to bind specifically other amino acids, while RNA and DNA can selectively interact through Watson-Crick base pairs [17]. Of course, proteins used in extant organisms are composed of 20 amino acids and have evolved into precise molecular machines that could not be produced without genetic information. Most researchers concluded that life could not emerge from proteins alone.

As an alternative, fifteen years ago I proposed the GADV hypothesis which was based on computer analysis of the databases of extant microbial genes and proteins. The hypothesis assumes that life emerged from a [GADV]-protein world, which was formed by pseudo-replication of [GADV]-protein in the absence of any genetic function [2,3,19]. I reasoned that [GADV]-protein could be pseudo-replicated if water-soluble globular and flexible [GADV]-proteins with a weak but sufficient activity for peptide bond formation could be produced by chance on the primitive Earth. Such a protein could then synthesize other water-soluble proteins having a similar globular structure by direct random joining of [GADV]-amino acids without genetic guidance (see also Sections 4.1 and 4.2).

At first, it is difficult to imagine how a small protein composed of one hundred of [GADV]-amino acids could be polymerized in prebiotic conditions. (The number of 100 is proposed here, because a small globular protein or a single domain protein is usually composed of around 100 amino acid residues [18]). Therefore, [GADV]-protein would actually be aggregates of [GADV]-peptides which assembled by non-covalent forces such as hydrophobic interactions and ionic bonds. The aggregate peptides could be formed as water-soluble globular structures if they were composed of roughly equal amounts of [GADV]-amino acids. Based on data from extant microbial genes encoding water-soluble globular proteins [20], these four amino acids are incorporated into secondary structures such as alpha helices, beta sheets, turn/coils and tertiary structures stabilized by hydrophobicity/hydrophilicity. To summarize, the earliest protein-like

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

molecules evolved from aggregates of short GADV peptides to GADV protein composed of about 100 amino acid residues. For simplicity in the rest of this review, I will refer to hypothetical water-soluble peptide aggregates as [GADV]-proteins.

We have experimentally confirmed that [GADV]-proteins in the form of aggregated [GADV]-peptides, can be synthesized by repeated heat-drying processing of [GADV]-amino acids. Furthermore, the polymers have protease activity as well as exhibiting the reverse reaction of forming peptide bonds [21]. Other researchers have confirmed that di- or tri-peptides exhibit catalytic activity [22,23] and that even peptides as simple as Gly-Gly, and Gly-Gly-Gly can catalyze peptide bond formation between chemically activated amino acids [24].

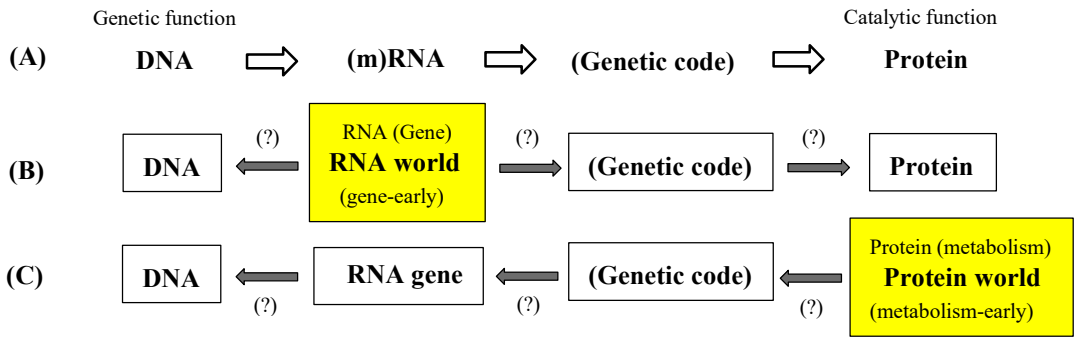

Figure 1. (A) Genetic information flows in the direction from DNA (genetic function) or (m)RNA to protein (catalytic function) in modern organisms (open arrows). The genetic system holds "chicken-egg relationship" between DNA (RNA) and protein, because genetic function on DNA cannot be expressed without protein and protein cannot be produced without gene (DNA). (B) In RNA world hypothesis, it is considered that formation of the first fundamental life system began from RNA (RNA world) with dual functions (genetic function and catalytic function), and later, genetic and catalytic functions were transferred to DNA and protein, respectively. (C) In protein world hypothesis as GADV hypothesis, the formation began from protein (protein world) and successively progressed to establish genetic code, RNA and DNA as going up from the lower (protein world) to the upper stream (gene) of the genetic flow in the present life system. Closed arrows and question marks means the steps still unidentified well.

The [GADV]-protein hypothesis is based on the sequence of events illustrated in Figure 1 C in which the first genetic system was established to synthesize increasing amounts of [GADV]-protein, ranging from globular aggregates of short GADV-rich peptides to more complete water-soluble globular proteins. In the rest of this review I will briefly describe the process by which I arrived at the set of four [GADV]-amino acids as components of primitive water-soluble globular proteins. This began with an analysis of seven microbial genomes in terms of the amino acid composition related to secondary and tertiary structures [25], then to a hypothesis related to a genetic code that involved ten amino acids [20,26], and finally to the GADV hypothesis presented here.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

The weak but sufficiently high catalytic activity of the protein led to the emergence of life. Therefore, the hypothesis is an idea based on "metabolism(protein)-early theory" (Figure 1 (C)). Thus, I consider that the first genetic system was established to synthesize more and more refined [GADV]-protein, from globular aggregate of short GADV-rich peptides to more complete water-soluble globular protein with a high molecular weight, step by step.

I briefly describe the process here, which I arrived at the set of four [GADV]-amino acids for producing the most primeval water-soluble globular protein. I actually but accidentally started from the study on creation of entirely new (EntNew) gene/protein or the first family gene/protein in microorganisms living on the present Earth, therefore, but not from that on the origin of life. From the results analyzed data bases of seven microbial genomes using six conditions ((the four condition (hydrophobicity/hydrophilicity, α-helix, β-sheet and turn/coil formabilities) plus acidic and basic amino acid compositions) for formation of water-soluble globular protein, I first arrived at GC-NSF(a) hypothesis for creation of EntNew gene/protein [25]. Successively, I tackled the study on the origin of genetic code and obtained GNC primeval genetic code hypothesis through SNS primitive genetic code hypothesis encoding 10 amino acids [20,26]. N and S mean four bases (A, U, G and C) and guanine, G, or cytosine, C, respectively. Finally, I hit upon the GADV hypothesis on the origin of life [2,3], from that GNC encodes four [GADV]-amino acids.

2. What is the Key point for solving the Origin of Life?

All organisms inhabit under the fundamental life system composed of gene, genetic code, and protein (metabolism) on the present Earth (Figure 1 (A)). Therefore, adequacy of a hypothesis on the origin of life should be judged by whether the formation processes of the three components, gene, genetic code and protein, can be reasonably explained by the hypothesis, or not.

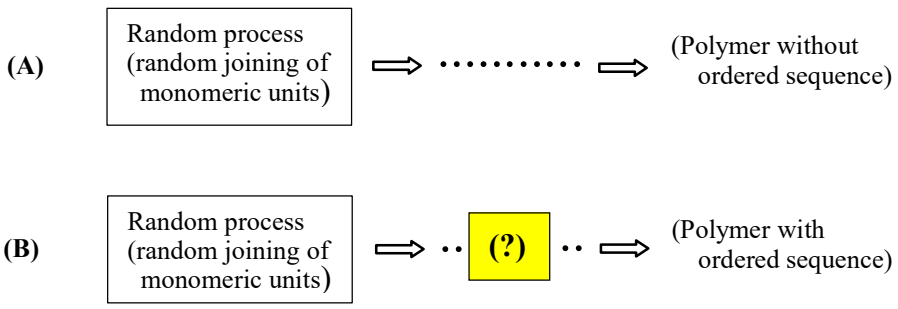

Figure 2. (A) Only essentially random process should occur on the primitive Earth. However, generally speaking, polymers with ordered sequence, such as protein and RNA, could not be produced by joining of monomeric units, because ordered sequence cannot be designed from the beginning. (B) As well known, organisms living on the present Earth have used many proteins and RNAs with ordered sequence. Therefore, it is supposed that something ingenious (indicated with question mark in yellow box) makes it possible to convert from random polymerization to the synthesis of polymer with ordered sequence. The main theme of this article is to know what made the conversion possible.

As a matter of course, it would be adequate to assume that only essentially random polymerization of monomeric units occurred on the primitive Earth, although a biased polymer

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

synthesis might be progressed due to intrinsic chemical reactivity among terminal residue and incoming ones. The reason is also because any spesific amino acid sequence could not be produced before establishment of the genetic system. Furthermore, it is generally supposed that only meaningless polypeptide and RNA should be produced by the random polymerization of the respective monomeric units (Figure 2 (A)). Contrary to that, many polymers with ordered sequence are produced in modern organisms under the genetic function (Figure 1 (A)).

Therefore, it is quite difficult to understand how the random polymerization could be converted to the synthesis of polymer with ordered sequence during repeated random processes. Inversely saying, it would be essential to know what caused the conversion from random to systematic polymer synthesis for elucidating the riddle on the origin of life (Figure 2 (B)) [27]. Taking these points into consideration, it is discussed in the following sections, which hypothesis, RNA world hypothesis and GADV hypothesis, can rationally explain the evolutionary process from accumulation of simple organic compounds on the primitive Earth to the emergence of life.

3. Can Establishment Process of the first Fundamental Life System be explained from the Standpoint of RNA World Hypothesis?

In this Section, I will discuss propriety of RNA world hypothesis on the question, in order of gene, genetic code and protein, according to the evolutionary process assumed by the hypothesis (Figure 1 (B)), one by one. Before that, I explain that there originally exist some weaknesses in the RNA world hypothesis. First, RNA is unstable in most chemical conditions. In addition, it would be quite difficult to synthesize both nucleotides and RNA under prebiotic conditions [2,3]. Furthermore, self-replication of RNA must be practically impossible due to the following self-contradiction. That is, RNA without stable tertiary structure would be required as a template to replicate RNA sequence and, simultaneously, RNA must be folded into a stable tertiary structure to exhibit its catalytic function [28].

3.1. Can RNA world hypothesis explain the formation process of the first gene?

It is considered according to the original RNA world hypothesis, that RNA world was formed upon accumulation of self-replicated RNAs, and the most primitive metabolic system was established in the world as various kinds of catalytic RNAs or ribozymes gradually accumulated on the primitive Earth. It is also considered that the catalytic RNAs acquired the most primitive genetic code and genetic information for protein synthesis (Figure 1 (B)), although not addressed explicitly in the hypothesis.

However, the first RNA must be produced by essentially random process, although, as a matter of course, the process was not purely random, because of a biased polymerization. The RNA, which was produced by random joining of mononucleotides, would be meaningless polynucleotide, even if RNA could be synthesized non-enzymatically on the primitive Earth (Figure 2 (A)). Therefore, the first RNA should code for mere random polypeptide because diversity of nucleotide sequence is extraordinary large as $(4^3)^{100} = \sim 10^{180}$ even for synthesis of a small protein composed of only 100 amino acids. The reason why the nucleotide sequence diversity ($\sim 10^{180}$) is much larger than amino acid sequence diversity ($20^{100} = \sim 10^{130}$) is simply because the universal or standard genetic code degenerates mainly at the third codon position. Therefore, the nucleotide sequence diversity

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

should coincide with the amino acid sequence diversity, if contribution of the degeneracy is disregarded.

In addition, self-replicated RNA also could not acquire any meaningful genetic information for protein synthesis, even if the first RNA produced by random polymerization could be repeatedly self-replicated (Table 1). The RNA should develop its function for self-replication and for its template in extraordinary large space to obtain a higher catalytic activity during repeated self-replication, but not to gain genetic information for protein synthesis. These indicate that the genetic information could not be encoded on the self-replicated RNA, because the two functions of the RNA is totally different. This suggests that mere random polypeptide would be produced, even if polypeptide synthesis was carried out under direction of the self-replicated RNA.

3.2. Can RNA world hypothesis explain the establishment process of the first genetic code?

It would be also quite difficult to propose what was the first genetic code and how it was created, from the standpoint of RNA world hypothesis (Table 1). The reason is because all combinations of triplet base sequence or $4^3 = 64$ kinds of triplets should be appeared on RNA strand if the RNA was produced by random polymerization of mononucleotides, independently of the number of amino acids used in the first genetic code. Note that the first primeval genetic code must be triplet from the beginning, but not singlet or doublet, otherwise, the meaning of all codons used would be lost, when the first code shifted to the present triplet code. This indicates that, in the case of RNA, which was produced by random polymerization of mononucleotides, 64 codons including termination codon(s) must be prepared for protein synthesis from the beginning. In order to avoid the difficulty upon the shift of singlet code to triplet code, Higgs has proposed four column theory for the first genetic code, which encodes four [GADV]-amino acids and has complete degeneracy at the first and the third codon positions [29]. However, I consider that the genetic code assumed by Higgs could not be realized on the primitive Earth, because wobble base pairing should require a much higher recognition mechanism on the most primitive ribosome. Therefore, the first genetic code must be read with only the Watson-Crick base pairing, which binds simply with each other at the most stable position.

3.3. RNA world hypothesis cannot explain the formation process of the first protein, too

Almost all researchers working in the field of the origin of life would consider that the first protein should be produced by transfer of catalytic activity on the self-replicated RNA or ribozyme onto a protein, as suggested by RNA world hypothesis. However, it would be actually impossible to directly transfer catalytic center on RNA with three-dimensional structure onto a three-dimensionally folded protein [2, 3, 28].

In addition, protein synthetic system or genetic system must be formed some day, separately from the transfer of the catalytic site from RNA to protein, even if the catalytic site on self-replicated RNA could be transferred to protein. However, it would be also impossible to establish the genetic system through fitting genetic code or codons to nucleotide sequence of self-replicated RNA. Needless to say, it would be also impossible to indirectly transfer catalytic center on RNA onto protein by using genetic information on the RNA, because probability that base sequence of the ribozyme codes for a protein with the same catalytic activity is essentially zero, judging from the extraordinary large diversity of RNA sequence ($(4^3)^{100} = \sim 10^{180}$), as described above. Therefore, it

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

cannot be also explained based on the RNA world hypothesis how the first protein could be produced (Table 1).

As described above, it would be obvious that the most important point for solving the riddle on the origin of life or the formation processes of all the three components, gene, genetic code and protein, cannot be explained with RNA world hypothesis. Therefore, I worry about that the study on the origin of life would be misled to a wrong direction during a long period further, if researchers continue to insist on the RNA world hypothesis from now on, because catalytic function of RNA or ribozyme has been overestimated in the hypothesis as a key for solving the riddle on the origin of life.

Of course, I do not assert that RNA did not play any role in the origin of life. Far from that, I understand that life could never emerge without RNA or gene. I would like to stress here that not RNA but [GADV]-protein, which was prepared on the primitive Earth, triggered the developmental process to establish the first genetic system and, consequently, life emerged at an evolutionary step, when enough number of RNA genes encoding protein could be obtained, as described in the following Sections.

Table 1. Evolutionary steps to establishment of the first genetic system presumed by RNA world hypothesis. Yellow boxes indicate evolutionary stage appeared newly. Question marks indicate the stages containing something unknown. Many question marks can be seen in the Table. It means that there are many steps, which cannot be explained or cannot be identified by RNA world hypothesis.

Stage	Organic catalyst	Gene	Genetic code	Protein
1. before RNA world	-	-	-	-
2. in RNA world	Ribozyme (RNA)	ss-RNA (?)	-	-
3. Genetic code	RNA or Protein (?)	ss- or ds-RNA (?)	(?)	-
4. Protein	Protein (?)	ds-RNA (?)	(?)	(?)

4. The Process until the first Genetic System was established, can be reasonably explained by [GADV]-protein World Hypothesis

Then, I will discuss whether the establishment process of the first genetic system can be reasonably explained by GADV hypothesis, or not, in order of protein, genetic code and gene, according to the evolutionary process assumed by the hypothesis (Figure 1 (C)).

4.1. Can GADV hypothesis explain the formation process of the first protein?

The first protein must be also produced by random joining of amino acids, which were accumulated on the primitive Earth. However, only useless polypeptide without ordered sequence should be produced by the random process, similarly to the case of RNA (Figure 2 (A)), because, as a matter of course, amino acid sequence of the protein is also never designed beforehand.

On the other hand, it is the unquestionable fact that the first life, which lived using biopolymers with ordered sequence as protein and RNA, emerged during repeated random processes. In addition, catalytic activity on proteins should be the most essential one for life.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

Therefore, protein with water-soluble globular structure must be synthesized at some point in order to exhibit a high catalytic function during the random joining of amino acids, which would be repeatedly carried out (Figure 2 (B)). What could overcome the seemingly impossible?

Protein 0^{th}-order structure played the key role in converting from random to ordered polymerization on the primitive Earth, because the polypetide chain produced by random joining of monomeric units in protein 0^{th}-order structure, actually by randomly polymerizing [GADV]-amino acids, could be folded into water-soluble globular structure necessary to exhibit a sufficiently high catalytic function (Figure 2(A), Figure 3) [30]. Peptides could be produced with [GADV]-amino acids accumulated in hot and salty environments on the primordial Earth, as shown with salt-induced peptide formation in a hot prebiotic ocean in the presence of metal ion as Cu^{2+} [31-33]. Furthermore, it is also shown by Luisi's group that de novo proteins obtained from a library of totally random polypeptides can be folded into a stable three-dimensional structure [34] and that long polymers can be formed by the so called 'fragment condensation' not by a stepwise polymerization [35].

The water-soluble globular [GADV]-protein or the aggregate of [GADV]-peptides, but not short peptides as di- and tri-peptides, could lead to formation of the genetic system for protein synthesis and to the emergence of life, although it is well known that short peptides, as di- and tri-peptides, have some catalytic activity [24]. The reason is because the genetic system is generally used for synthesis of a protein carrying more than 100 amino acids but not of short peptide. In other words, formation of water-soluble globular aggregate or protein and its catalytic activity hold the key for understanding the establishment of the genetic system and the origin of life.

4.2. How could protein 0^{th}-order structure convert from random to ordered polymerization?

Protein 0^{th}-order structure is a specific amino acid composition, in which even random joining of amino acids can produce water-soluble globular protein with a slightly more flexible structure than extant or mature protein, at a high probability (Figure 3) [30]. The phrase, "a high probability", is used in the meaning of "not always" here, because the amino acid composition of a protein synthesized with random joining of amino acids even in protein 0^{th}-order structure does "not always" satisfy the conditions for water-soluble globular protein formation, judging from simply a viewpoint of probability. A pool containing roughly equal amounts of [GADV]-amino acids or the specific amino acid composition is one of the protein 0^{th}-order structures (Figure 3).

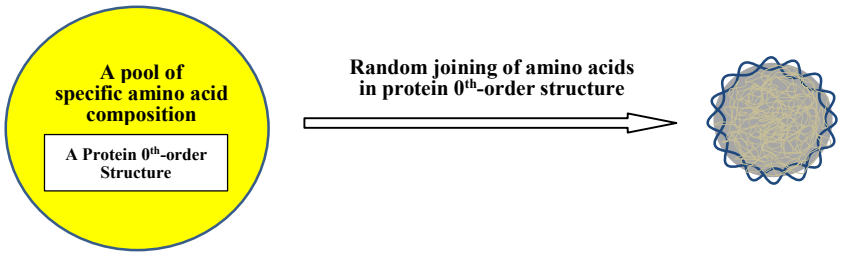

Figure 3. Protein 0^{th}-order structure holds the key for solving the riddle on the origin of life. That is, protein 0^{th}-order structure or a specific amino acid composition, in which random joining of amino acids (a process generally producing only useless one) produces water-soluble globular protein (useful one), made it possible to convert from random process to synthesis of protein with ordered sequence. Wavy lines on the surface of and thin curves in the circle indicate flexibly wobbling structure of the protein.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

Of course, such a pool containing equal amounts of [GADV]-amino acids, even roughly, would not exist anywhere on the primitive Earth. Even though, water-soluble globular [GADV]-protein could be produced on the primitive Earth, as described in more detail in Section 4.4. The reason, why the [GADV]-amino acid composition is a protein 0^{th}-order structure, is described in the next Section. Inversely saying, water-soluble globular [GADV]-protein could be produced by direct random joining of amino acids "at a high probability", owing to the protein synthesis in 0^{th}-order structure.

Another reason, why water-soluble globular [GADV]-protein could be produced by direct random joining of amino acids, is because individual amino acids are a functional unit having specific characteristics, as hydrophobicity/hydrophilicity, α-helix, β-sheet and turn/coil formabilities, which are necessary to form water-soluble globular protein leading to acquisition of a sufficient catalytic activity (Table 2). The characteristics of individual amino acids makes it possible to produce catalytic protein even by random joining of [GADV]-amino acids, unlike the case of RNA, which exhibits genetic function with not one but three bases. Thus, the water-soluble globular proteins, which were synthesized and accumulated by random polymerization of the amino acids, could trigger the establishment of the first genetic code followed by creation of the first gene and the emergence of the first life (Figure 1 (C)).

4.3. The reason why [GADV]-amino acid composition is one of protein 0^{th}-order structures

Then, I explain the reason why [GADV]-amino acid composition is one of protein 0^{th}-order structures. First, structure index of extant protein from seven microorganisms was calculated with 20 amino acid compositions in the protein and structure index of the corresponding amino acid, according to the following equation; $I_p = \Sigma_i\ (Amino\ acid\ composition)_i \times (I_a)_i$ (Equation 1). I_p and I_a mean protein structure index and index of amino acid for secondary or tertiary structure formation, respectively [20,36]. The equation indicates that all proteins produced by random joining of amino acids in an amino acid composition have actually the same protein structure index or the same propensity forming secondary and tertiary structures of the protein, as a whole, because all the proteins have basically the same amino acid composition and, therefore, the same protein structure index, although amino acid sequence of the proteins is quite different from each other. Of course, amino acid sequence of protein, rather than average amino acid composition, dominates the final property of being water soluble, because whole protein structure is controlled by local secondary structures at the level of 3-10 amino acids, and interactions between these small structures determine a stable three-dimensional structure. However, I discuss formation of water-soluble globular protein as a whole, which is determined by amino acid composition, and that the immature protein with some flexibility, which is produced by random joining of amino acids in protein 0^{th}-order structure is important to understand the establishment process of the genetic system and the conversion from immature protein produced by random process to mature protein with ordered structure, in this article (Figure 3).

From the results analyzed with the equation 1, it was found that four protein structure indexes are roughly constant, even if a half number of amino acid compositions of extant protein varied largely upon change of GC content of a gene [20]. Thus, I concluded that the four structure indexes of protein can be used to judge whether a polypeptide chain is folded into water-soluble globular structure, which is pre-requisite to exhibit a sufficiently high enzymatic function, or not.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

From the analysis with the four indexes, it was found that [GADV]-polypeptide chain could be folded into water-soluble structure at a high probability, when the polypeptide contains roughly equal amounts of [GADV]-amino acids (Table 2). This indicates that even a polypeptide chain could be folded into water-soluble globular structure at a high probability, if the polypeptide was synthesized by random joining of the [GADV]-amino acids in the protein 0^{th}-order structure in the absence of genetic function (Figure 3).

Optimum contents of [GADV]-amino acids are shown in Table 2, as average values when [GADV]-protein satisfied the four conditions for water-soluble globular protein structure formation. Of course, [GADV]-protein produced by random joining of amino acids on the primitive Earth was quite incomplete and did exhibit only a quite low catalytic activity, because a pool containing equal amounts of [GADV]-amino acids did not exist anywhere on the primitive Earth. However, it can be supposed that even the incomplete [GADV]-proteins would played sufficient roles in catalyzing chemical reactions without any proteinaceous catalyst on the primitive Earth, as described in more detail in the following Section 4.4. The second and third protein 0^{th}-order structure are SNS-encoding 10 amino acids and amino acid composition encoded by antisense codon sequence of GC-rich gene (GC-NSF(a)).

Table 2. Characteristics of [GADV]-amino acids. Optimum content of [GADV]-amino acids for producing water-soluble globular protein were obtained with computer analysis as average values, when computer-generated amino acid compositions satisfied the four conditions (hydropathy, α-helix, β-sheet and turn/coil formabilities) for protein structure formation. The results of optimum contents of [GADV]-amino acids indicate that [GADV]-protein have similar rigidity to extant protein, when [GADV]-proteins contain less and more quantities of glycine and alanine than one quarter, respectively. Therefore, it is supposed that [GADV]-proteins are more flexible than the presently existing proteins, when the proteins contain [GADV]-amino acids at one quarter each.

Amino acid	One L.S.[1)]	Codon	Optimum C.[1)]	α-Helix [2)]	β-Sheet [2)]	Turn/coil [2)]	Hydrophob. [1) 2)]
Glycine	[G]	GGC	0.111	0.56	0.92	1.64	1
Alanine	[A]	GCC	0.313	1.29	0.9	0.78	1.6
Aspartic acid	[D]	GAC	0.289	1.04	0.72	1.41	-9.2
Valine	[V]	GUC	0.287	0.91	1.49	0.47	2.6

[1)] One L.S., Optimum C. and Hydrophob. are abbreviated terms of "one letter symbol", "optimum content" and "hydrophobicity", respectively. [2)] The number in yellow box means that the amino acid have the index for propensity of the respective secondary or tertiary structures in top-five out of 20 natural amino acids.

4.4. Can GADV hypothesis explain the formation process of the first protein?

Water-soluble globular [GADV]-protein could be produced by direct random joining of [GADV]-amino acids in a pool containing roughly equal amounts of [GADV]-amino acids, as described above. However, such a pool would not actually exist on the primitive Earth, because

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

[GADV]-amino acids should be accumulated on the primitive Earth in order of Gly, Ala, Asp and Val. In addition to that, many various kinds of organic compounds other than amino acids should be also contained in the pool, such as depressions of rocks on seashore of the primitive Earth. However, positive and negative charges on amino and carboxyl groups could help to pull against amino acids each other, and a strong hydrophobic interaction among valine residues could assist to collect valine-containing [GADV]-peptides in water. Thus, the biased accumulation of [GADV]-amino acids on the primitive Earth could be relatively corrected in aggregate of [GADV]-peptides.

Nevertheless, [GADV]-protein formed by direct random joining of [GADV]-amino acids was quite incomplete because of still largely biased [GADV]-amino acid composition of the protein. In addition, the protein in the world was actually aggregate of [GADV]-peptides. Even such incomplete [GADV]-protein could exhibit weak but enough catalytic activity on the surface of the protein, because water-soluble globular structure with some flexibility could adapt to a substrate to exhibit the activity (Figure 3).

4.5. Can GADV hypothesis explain the establishment of the first genetic code?

Genetic code connects genetic function on DNA or RNA with protein synthesis or actually triplet nucleotide sequence or codon with amino acid (Figure 1 (A)). In addition, the genetic code specifies a framework of amino acids, which can be used in protein synthesis. Of course, even polypeptide chain, which was synthesized by indirect random joining of amino acids under the first genetic code, must be folded into water-soluble globular structure (Figure 3). Otherwise, the genetic code could not play the role in synthesizing protein with catalytic activity. Taking them into consideration, the first genetic code must encode amino acids in the simplest protein 0^{th}-order structure.

Furthermore, it would be favorable for the first genetic code to encode a small number, four or so, of amino acids, because it should become much more difficult to establish the genetic code, as the number of amino acids becomes larger. Moreover, the first genetic code must encode all amino acids in the simplest protein 0^{th}-order structure. The reason is because the genetic code should be established, when all amino acids in the protein 0^{th}-order structure were incorporated into the genetic code.

We have previously proposed GNC-SNS primitive genetic code hypothesis, suggesting that the universal or standard genetic code originated from GNC code encoding [GADV]-amino acids or the first protein 0^{th}-order structure through SNS code encoding 10 amino acids in the second protein 0^{th}-order structure [20]. Therefore, essentially random polymerization of amino acids were carried out under the GNC code, after the first genetic code was established. The polypeptide chain produced could be folded into water-soluble globular structure at a high probability, because random joining of amino acids through GNC code is the same as [GADV]-protein synthesis by direct random joining of [GADV]-amino acids.

The establishment of GNC genetic code made it possible to use only [GADV]-amino acids for protein synthesis for the first time. The [GADV]-amino acid sequence space (4^{100} = ~10^{60}) encoded by the GNC code is much smaller than the full space (20^{100} = ~10^{130}) composed of 20 amino acids. In addition, foldable polypeptide sequences are contained in the [GADV]-amino acid sequence space at a much higher ratio than the full space. The reason is because the full space contains an

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

extraordinary large number of non-foldable sequences encoded by AT-rich base sequence in addition to GC-NSF(a)-encoding sequence space (the third protein 0^{th}-order structure), which is similar to the space (the second protein 0^{th}-order structure) encoded by SNS genetic code (10^{100}). These could also help to produce water-soluble [GADV]-protein more efficiently than the full sequence space. Thus, many useful water-soluble globular proteins with a catalytic function could be produced by the indirect random joining of [GADV]-amino acids under the GNC code.

The establishment of the first GNC genetic code was accelerated by oligonucleotide synthesis in GADV-protein world. That is, the complex formation between one of [GADV]-amino acids and the corresponding GNC anticodon-containing oligonucleotide (the most primitive tRNA) made it possible to synthesize random but more refined [GADV]-protein under the GNC code. The primitive tRNA come from oligonucleotide synthesized in [GADV]-protein world as described in Section 5, stage 2. The tRNA could carry [GADV]-amino acids selected out by stereospecific interaction between one of [GADV]-amino acids and the corresponding GNC anticodon-containing oligonucleotide. At that time, arrangement of GNC codons in the primitive tRNA, side by side, could play the role in a primitive mRNA through tRNA dimer formation, as Giuimaraes has similarly considered [37]. The dimer formation of two tRNA on primeval ribosome or complex of [GADV]-protein(s) and oligonucleotide(s) or primeval rRNA(s) could help catalyzing peptide bond formation between two amino acids in the complexes through the effect adjoining two amino acids. At the present time, I assume that catalytic reaction for the peptide bond formation was carried out with [GADV]-protein in the primeval ribosome (see also Section 7.1.1. "Is peptidyl transferase activity on rRNA a vestige of RNA world?").

4.6. Can GADV hypothesis explain the formation process of the first gene?

Gene is surely composed of a number of mononucleotides. As described in Section 2, the first gene also must be created with a random process, because nucleotide sequence is never designed previously. On the other hand, it would be impossible to create gene carrying genetic information for protein synthesis by random joining mononucleotides, because nucleotide sequence diversity is extraordinarily large ($(4^3)^{100}$ = ~10^{180}), as described in Section 3.1. Here, it is important to note that gene is substantially not mononucleotide but codon sequence. Taking them into consideration, it can be concluded that the first gene must be created by random joining of GNC codons encoding [GADV]-amino acids. Therefore, water-soluble globular [GADV]-protein could be produced, if random GNC sequence or $(GNC)_n$ was the first gene.

After creation of the single-stranded $(GNC)_n$ gene, plural number of [GADV]-proteins with the same amino acid sequence could be repeatedly synthesized by expression of the $(GNC)_n$ gene, although the [GADV]-protein was only immature protein with random [GADV]-amino acid sequence. In order to create the first double-stranded $(GNC)_n$ gene, many difficulties must be overcome, such as establishment of the first GNC genetic code, the first single-stranded $(GNC)_n$ gene formation by joining of GNC codon in oligonucleotides or primitive tRNAs and complementary strand synthesis of the single-stranded gene to produce the first double-stranded $(GNC)_n$ gene, as seen in Table 3 and Fig. 5. However, such difficulties would not be problematic for the creation of the first double-stranded gene, because the previous situation remained, even if the difficulties could not be overcome at a step, and it could be repeatedly tried until the difficulties were finally overcome.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

The first double-stranded $(GNC)_n$ gene was created by complementary strand synthesis of the first single-stranded gene. The formation of the first double-stranded $(GNC)_n$ gene made it possible to evolve immature and rather flexible protein with a weak catalytic activity to mature rigid protein with a high catalytic activity, for the first time (Figure 4). In addition, first, evolvability was acquired by the formation of the first double-stranded gene (Table 3). Thus, the first genetic system was established, as three components were piled up in order of protein, genetic code and gene, one by one (Table 3). The conversion from random polymerization to the synthesis of protein with ordered sequence took place upon the formation of the first genetic system and the formation always progressed under one of protein 0^{th}-order structures or [GADV]-amino acids.

As is easily expected, it would also take a long time until the second $(GNC)_n$ gene was formed, if the gene were created *de novo* by random joining of GNC codons similarly as the case of the first gene (Figure 4). On the other hand, the first life could not live with only one gene encoding one protein, because life is a high-level concept associated to systems of molecules, cellularity, self-maintenance, reproduction and so on, probably all these features are linked to protocellular systems rather than few replicating molecules. However, many genes fortunately could be produced by using the first double-stranded $(GNC)_n$ gene, because both base sequences on sense strand of the gene and the corresponding antisense sequence are GNC codon sequences [30]. Therefore, both $(GNC)_n$ gene and its antisense $(GNC)_n$ sequence could code for [GADV]-proteins with similar amino acid composition but quite different amino acid sequence each other. This makes it possible to produce homologous protein in a protein family and EntNew protein or the first family protein from sense and antisense sequences of the first double-stranded $(GNC)_n$ gene, respectively, after gene duplication (Figure 4) [30]. Thus, life could emerge on the primitive Earth, because formation of the first double-stranded $(GNC)_n$ gene triggered creation of many kinds of genes encoding [GADV]-protein with a catalytic or a structural functions. Furthermore, new homologous and entirely new genes/proteins could be derived also from EntNew double-stranded $(GNC)_n$ genes originated from antisense codon sequence of the first gene.

Table 3. Evolutionary stages to establish the first genetic system presumed by GADV hypothesis, which I have proposed. Yellow boxes indicate evolutionary stages appeared newly. The table clearly shows that every stage is intimately related to [GADV]-protein, suggesting that the evolutionary stages evolved to produce [GADV]-protein with higher catalytic function than the protein used at one stage before. Single-stranded $(GNC)_n$ RNA gene was invented to synthesize plural number of [GADV]-proteins with the same catalytic function, and double-stranded gene was formed to make it possible to evolve protein function through introduction of base substitutions. Furthermore, the table also shows that all the evolutionary steps can be rationally explained by GADV hypothesis.

Stage	Protein	Genetic code	ss-Gene (mRNA)	ds-Gene
1. before [GADV]-protein world	-	-	-	-
2. in [GADV]-protein world	[GADV]-protein	-	-	-
3. GNC primeval genetic code	[GADV]-protein	GNC genetic code	-	-
4. Single-stranded $(GNC)_n$ gene	[GADV]-protein	GNC genetic code	ss-$(GNC)_n$ gene	-
5. Double-stranded $(GNC)_n$ gene	[GADV]-protein	GNC genetic code	$(GNC)_n$ mRNA	ds-$(GNC)_n$ gene

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

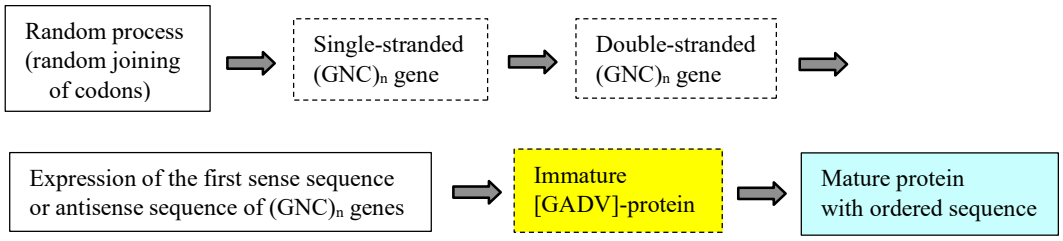

Figure 4. Both the first single-stranded and the first double-stranded $(GNC)_n$ genes were created through random process, because the double-stranded gene was formed by complementary strand synthesis of the single-stranded $(GNC)_n$ gene, which was produced by random joining of GNC codons. However, the formation of the double-stranded gene made it possible to generate mature protein with ordered sequence from sense sequence and to form entirely new gene/protein with ordered sequence from antisense codon sequence for the first time, through introduction of base replacements required onto the sense and antisense sequences, respectively.

As described above, all three components of the first genetic system, [GADV]-protein, GNC genetic code encoding [GADV]-amino acids, $(GNC)_n$ gene encoding [GADV]-protein, could be formed about 3.8 billion years ago, owing to [GADV]-amino acids or the protein 0^{th}-order structure (Table 3). Therefore, I consider that life might not emerge without the protein 0^{th}-order structure or [GADV]-amino acids on the primitive Earth.

5. Establishment process of the first genetic system assumed from GADV hypothesis

The establishment process of the first genetic system, which assumed by GADV hypothesis [2,3], is described step by step in this Section, so that the propriety of the hypothesis can be further confirmed by the description.

Stage 1: Before formation of [GADV]-protein world

Various kinds of organic compounds, as amino acids, amines, organic acids and so on, were synthesized through lightning in primitive atmosphere [38], at hydrothermal vents in deep sea [4-6] and *etc.* and accumulated on the primitive Earth in the absence of proteinaceous catalysts. The organic compounds would be also delivered to the early Earth by meteorites, asteroids, comets, and interplanetary dust particles (Table 4) [39].

Stage 2: After formation of [GADV]-protein world or in [GADV]-protein world

[GADV]-protein could be synthesized such as by repeated heat-drying process in depressions on rocks [21]. Successively, [GADV]-protein world was formed, as [GADV]-proteins were produced and accumulated by pseudo-replication of [GADV]-protein. Pseudo-replication is the synthesis of water-soluble globular [GADV]-proteins with not the same amino acid sequence but similar amino acid composition through direct random joining of [GADV]-amino acids. Only random physical, physicochemical and chemical reactions among various compounds, which were accumulated on the primitive Earth, progressed at Stage 1. The reason, why the pseudo-replication could be carried out even in the absence of any genetic function, is because [GADV]-amino acid composition is one of protein 0^{th}-order structures, as described in Section 4.3 (Figure 3). Thus, various organic compounds such as amino acids, peptides, nucleotides and oligonucleotides, could be synthesized with [GADV]-proteins in the [GADV]-protein world. Of course, it does not exist anything like a purely "random" oligo- or polymerization, because a biased polymerization would be carried out

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

due to intrinsic chemical reactivity among terminal residue and incoming ones. However, not purely but substantially random chemical reactions proceeded among various organic compounds at this stage, too, in the meaning of that systematic chemical reaction could not be carried out in the absence of genetic system, which instructs the synthesis of protein with a specific sequence.

[GADV]-proteins produced by the pseudo-replication inevitably contained amines, carboxylic acids and non-natural amino acids, which were accumulated on the primitive Earth at Stage 1 (Figure 5, Stage 2). However, the [GADV]-protein synthesized at Stage 2 might lead to preferential synthesis of oligonucleotides, because the fact is known that [GADV]-amino acids can be seen at active sites of extant proteins catalyzing RNA metabolism at a high frequency [40]. The preferential oligonucleotide synthesis with [GADV]-protein might accelerate the progress of evolutionary stages to the emergence of life through establishment of GNC primeval genetic code, as described following Stages.

Stage 3: After establishment of GNC genetic code

Protein synthesis in extant organisms is carried out mainly at two stages. One is amino acylation of tRNA between amino acid and tRNA. The other is peptide bond formation between amino acids on two aminoacyl-tRNAs. In this respect, I presume that the the former was catalyzed by [GADV]-protein with catalytic activity connecting amino acid with primeval tRNA and the latter was mediated by [GADV]-protein in primeval ribosome composed of primeval rRNA(s) and ribosomal protein(s) (see also Section 7.1.1).

Table 4. Evolutionary stages deduced from GADV hypothesis. It is considered that the evolutionary stage progressed to produce more useful and more complete [GADV]-protein than before, step by step. Yellow boxes indicate that the stage was newly formed. "[GADV]-protein" written in the column "Organic catalyst" is different from each other. That is, various kinds of organic compounds other than [GADV]-amino acids were contained in the protein synthesized at Stage 2. Both [GADV]-proteins produced at Stages 3 and 4, are, similarly, immature proteins with substantially random amino acid sequence. However, plural number of [GADV]-proteins with the same amino acid sequence could be synthesized at Stage 4. [GADV]-protein synthesized from sense sequence or gene at Stage 5 could be evolved from immature to mature protein, because both sequences of double-stranded $(GNC)_n$ gene could play the role of template in replicating the sequences each other, and base substitutions necessary to evolve could be preserved using antisense sequence as template..

Stage	Organic catalyst	Monomer [1)]	Randomness	Evolvability
1. before [GADV]-protein world	None	Various compounds	Random	-
2. in [GADV]-protein world	[GADV]-protein	Various compounds	Random	-
3. GNC primeval genetic code	[GADV]-protein	[GADV]-amino acids	Random	-
4. Single-stranded $(GNC)_n$ gene	[GADV]-protein	[GADV]-amino acids	Ordered sequence	-
5. Double-stranded $(GNC)_n$ gene	[GADV]-protein	[GADV]-amino acids	Ordered sequence	+

[1)] The term, "Monomer", means monomer used for polymer synthesis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

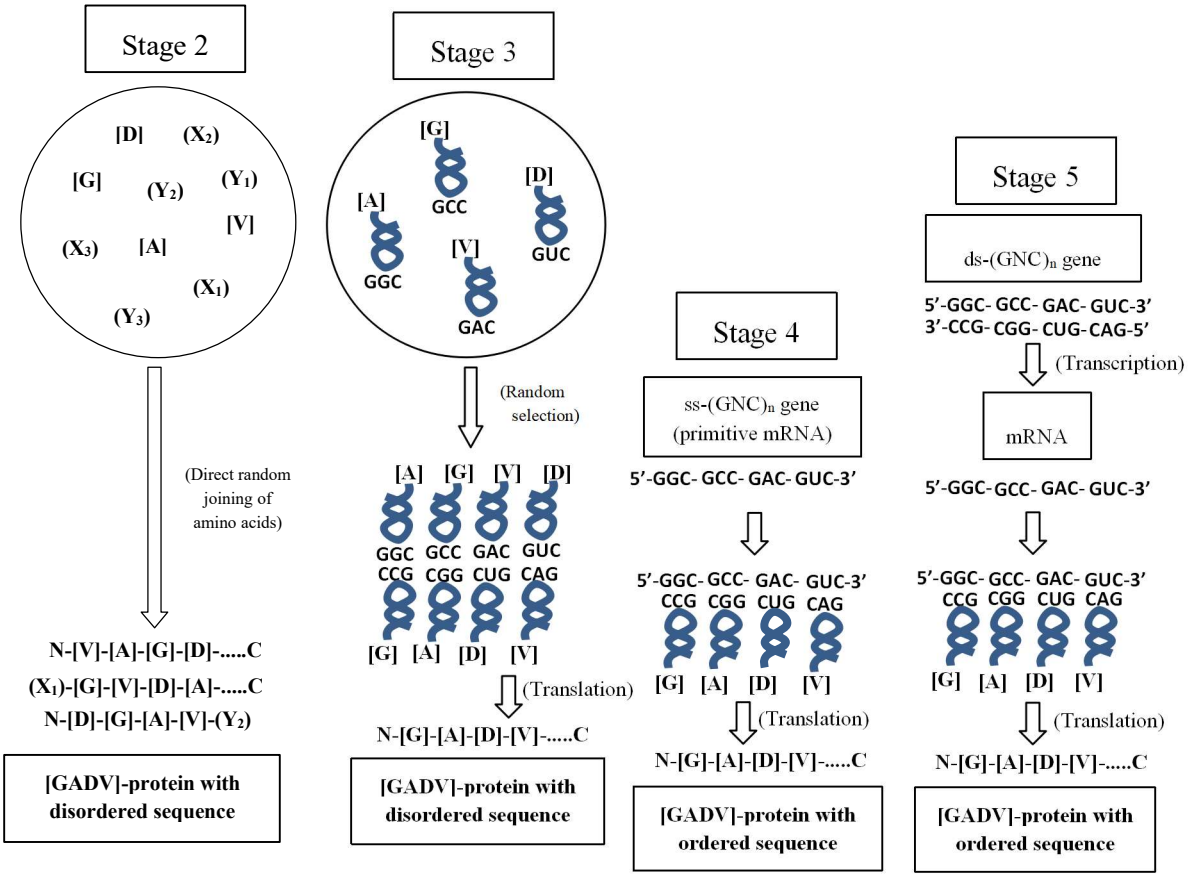

Figure 5. Evolutionary stages assumed by GADV hypothesis. Stage 2: [GADV]-protein, actually aggregates of [GADV]-peptides was produced by random joining of [GADV]-amino acids, which were accumulated at Stage 1 (not shown in this figure). [GADV]-protein produced at Stage 2 inevitably contained organic compounds as carboxylic acids (X_n) and amines (Y_n) other than [GADV]-amino acids. Stage 3: Random synthesis of [GADV]-protein was carried out through complexes of [GADV]-amino acid and oligonucleotide containing the corresponding GNC anticodon (the most primitive tRNA). The complex played the mRNA-like role in the protein synthesis each other. Synthesis of protein containing only [GADV]-amino acids could be carried out at Stage 3, for the first time. Stage 4: [GADV]-protein synthesis was carried out with single-stranded $(GNC)_n$ gene or primitive mRNA, which was produced by random joining of GNC anticodon contained in the complex of [GADV]-amino acid with the primitive tRNA. By using of single-stranded gene it became possible to synthesize plural number of [GADV]-proteins with the same catalytic activity. Stage 5: At this stage, transcription of double-stranded $(GNC)_n$ gene and successive translation of mRNA was carried out, upon creation of the double-stranded gene by synthesis of complementary strand of the single-stranded gene. Acquisition of the double-stranded gene made it possible to evolve immature [GADV]-protein to mature protein and to produce diverse new homologous genes/proteins and entirely new genes/proteins from sense and antisence sequences, respectively. ss- and ds- mean single-stranded and double-stranded, respectively. Random [GADV]-peptides after Stage 3 are drawn as [G]-[A]-[D]-[V]- only for simplicity. N and C in the amino acid sequences are described to show N-terminal and C-terminal ends of [GADV]-peptide, respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

Thus, the first GNC genetic code could be established through stereospecific interaction between one of [GADV]-amino acids and oligonucleotide containing GNC anticodon of the respective [GADV]-amino acids. Therefore, the GNC-containing oligonucleotide corresponds to the most primitive tRNA [41,42]. Side-by-side dimer formation between the two complexes could stimulate [GADV]-protein synthesis (Figure 5 (B)). The reason, why amino acids bound to the primitive tRNA were better linked together than reaction in the absence of the tRNA, is because of the effect adjoining two amino acids side by side through the two complexes. The complexes could also form another type of dimer, one complex opposite another complex through base pairing between two GNCs in the complexes [37]. Formation of the latter dimer could assist to place another dimer side by side (Figure 5; Stage 3). At the Stage 3, [GADV]-protein, which was synthesized under the GNC code, contained only [GADV]-amino acids. However, essentially random [GADV]-protein synthesis still continued at Stage 3 (Table 4).

Stage 4: After single-stranded $(GNC)_n$ gene formation

Single-stranded $(GNC)_n$ RNA gene would be formed by random joining of GNC codons in the complex of [GADV]-amino acid with GNC-containing oligonucleotide. The protein 0^{th}-order structure was also essential to the formation of the first RNA gene, because GNC is one of the first genetic code encoding [GADV]-amino acids (Figure 5; Stage 4). This means that RNA gene could never be formed in advance of establishment of the first genetic code.

Hereafter, synthesis of plural number of [GADV]-proteins or [GADV]-peptides with the same amino acid sequence was made possible, for the first time. The role of single-stranded $(GNC)_n$ gene in protein synthesis corresponded to mRNA in modern genetic system (Table 4, Figure 5; Stage 4).

Stage 5: After double-stranded $(GNC)_n$ gene formation

Double-stranded $(GNC)_n$ RNA gene was produced through complementary strand synthesis of the first single-stranded $(GNC)_n$ gene. At this stage, it was, first, made possible to evolve from immature, flexible [GADV]-protein with a weak catalytic activity to the mature, rigid protein with a high catalytic activity through introduction of required base substitutions onto the sense sequence (Figure 4, Table 4).

In parallel, conservation and propagation to progenies of genetic information on $(GNC)_n$ sequence became possible through gene duplication at Stage 5. Furthermore, creation of homologous and EntNew genes also became possible [30]. After generation of plural number of double-stranded $(GNC)_n$ RNA genes, the first life would emerge on the primitive Earth in due time. Thus, it goes without saying that RNA played the key role in the emergence of life.

I would like to emphasize here again that it is essential to take protein 0^{th}-order structure into consideration for understanding the conversion from random joining of amino acids to the synthesis of protein with ordered sequence and the emergence of life [30]. Furthermore, I have considered that the so-called "chicken-egg relationship" between protein and gene was formed as going up the modern genetic flow counter directionally, as assumed by GADV hypothesis. (Figure 1 (C), Table 3) [2,3,30].

7. Discussion

It has been widely considered that gene or RNA must be created before protein synthesis and even before the first genetic code on the primitive Earth. The reason is because it has been considered that protein could not be produced without gene, and that protein, itself, could not be replicated, as described in Introduction.

 doi:10.20944/preprints201712.0170.v1

However, I consider that evolutionary processes assumed by "gene-early theory" or "RNA world hypothesis" could not be realized on the primitive Earth, even if RNA was produced by random joining of mononucleotides (although it would be still questionable whether RNA could be really synthesized with prebiotic means). The reason is because there exist several weak points as described in Section 3. Such as, (1) Chemical instability. (2) Difficulties of nucleotide and RNA syntheses. (3) Difficulty of RNA self-replication. (4) Random RNA cannot encode a water-soluble protein with catalytic function. (5) RNA world hypothesis cannot explain the formation process of the genetic system composed of gene, genetic code and protein.

Therefore, I would like to point out that the origin of life cannot be explained from the standpoint of RNA world hypothesis, which depends on gene(RNA)-early theory. Nevertheless, many studies have been carried out to overcome the weaknesses of RNA world hypothesis. In fact, two interesting papers were published recently, one is about non-enzymatic synthesis of RNA by Szostak group [43] and another is on RNA replication with artificial ribozyme by Joyce group [44]. However, RNA primer and RNA template, which were previously prepared, were used in both studies. Therefore, it might newly become problematic to explain how RNA primer and RNA template were produced with prebiotic means. Therefore, it would be doubtful whether RNA could be really non-enzymatically synthesized as supposed by Szostak [43], and also RNA could be actually replicated on the primitive Earth, as supposed by Joyce [44].

However, RNA hypothesis has been accepted by many persons including researchers working in the research field of the origin of life still now. The reasons might be because many persons have considered the following facts as grounds supporting the hypothesis. Then, I discuss the two facts, that peptidyl transferase activity is on rRNA and that nucleotide-related compounds are used as cofactors in extant enzymes, as followings.

7.1.1. Is peptidyl transferase activity on rRNA a vestige of RNA world?

It is well known that peptidyl transferase activity positions at not ribosomal protein but rRNA in ribosome [45,46]. In RNA world hypothesis, it is considered that almost all ribozyme activities used in the world were transferred to the respective proteins. However, catalytic site of ribozyme with three-dimensional structure could not be, in principle, transferred onto protein with independently folded tertiary structure, as described in Section 3.3. On the other hand, it has been hypothesized that the peptidyl transferase activity, which was used for protein synthesis in RNA world, has not been transferred to protein, and, therefore, the activity remains still now on rRNA as a vestige of the RNA world. Only one exception, where catalytic activity can be transferred from RNA to protein or inversely from protein to RNA, should be the case, in which RNA and protein stand close to each other as like in ribosomal particle. Therefore, it should be understood as that the reason, why the peptidyl transferase activity exists on not protein but rRNA, is rather because the activity, which originally located on one of ribosomal proteins, was later transferred to rRNA, and but not because the activity remains on rRNA as a vestige of RNA world.

It is also well known that various ribozymes have been used in many organisms at the present time. However, catalytic activities of natural ribozymes are restricted in a small number of reactions such as scission and exchange reactions of phosphodiester bond and peptide bond [47,48]. In addition, catalytic activities of ribozymes are not used for reactions on major metabolic pathways. Therefore, it should be considered that those ribozymes were created not in RNA world but in rather recent eras.

7.1.2. Are nucleotide-related cofactors of modern enzymes vestiges of RNA world?

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2017 doi:10.20944/preprints201712.0170.v1

It is also well known that nucleobases can be synthesized by Miller's type experiments [38] and those are contained in carbonaceous meteorites from space. From these facts, it is frequently advocated that the nucleotide-related cofactors also should be vestiges of the RNA world [49,50]. However, the nucleotide-related compounds, which were synthesized by Miller's type experiments and detected in meteorites, are not nucleotides but nucleobases and their derivatives, suggesting that nucleobases but not nucleotides were used as cofactors on the primitive Earth. In fact, catalytic point of nucleotide-related cofactors is generally located on the nucleobase and their derivatives but not on ribose and phosphate moieties of nucleotides. Therefore, it would be reasonable to consider that ribose and phosphate were added to nucleobase to discriminate nucleotide-related cofactors to from other organic compounds more precisely than before.

P. Higgs, editor of the special issue, "The RNA World and the Origin of Life", describes that *"the RNA World has become a widely-accepted hypothesis for the Origin of Life, and has reached the level of a 'textbook theory'. Nevertheless, many issues related to the RNA World remain to be demonstrated fully in experiments, and important theoretical questions remain as to how RNA-based life could have originated, survived, and evolved."*

This also clearly indicates that the riddle on the origin of life has not been solved well with RNA world hypothesis even now, after about 30 years has already passed since the hypothesis was proposed by W. Gilbert [1]. The reason, why the origin of life has not been solved as yet, would be mainly because many persons has insisted on the hypothesis without understanding the significance of protein 0^{th}-order structure in the establishment process of the first genetic system. Therefore, it seems to me that the time is approaching when novel idea based on a new concept other than RNA world hypothesis should be taken into consideration in order to solve the riddle on the origin of life. I would like to assert that the novel idea is [GADV]-hypothesis based on the new concept; protein 0^{th}-order structure, both of which I have proposed [2,3,30].

Life emerged according to metabolism(protein)-early theory or GADV hypothesis!?

I believe that life emerged from [GADV]-protein world, because GADV hypothesis can explain reasonably the genetic system composed of gene, genetic code and protein, as described in this review article. Of course, it should not be concluded that life emerged on the primitive Earth as an idea assumes, even if the evolutionary process to the emergence of life can be explained by the idea, rationally or without large contradictions. However, inversely, if the evolutionary process cannot be reasonably explained with an idea, it could be concluded that the idea is wrong. Therefore, I would like to say that life did not emerge from "RNA world".

At the end of this article, I describe the weakness of GADV hypothesis, because I would like to contribute to development of the study on the origin of life based on the hypothesis. That is, we have only a small number of experimental results; such as, protease and RNase activities of [GADV]-proteins, actually aggregates of [GADV]-peptides, which were produced by repeated heat-drying processes [21]. One of the reasons is because we have carried out the study on GADV hypothesis mainly with computer analyses about the databases of extant microbial genes and proteins.

Therefore, I consider that the following steps should be confirmed by experiments to overcome the weakness of GADV hypothesis as described above and to solve the riddle on the emergence of life according to the hypothesis.

(1) Can [GADV]-polypeptide containing at roughly equal amounts of [GADV]-amino acids, which was synthesized with random joining of amino acids, be folded into water-soluble globular structure?
(2) Can water-soluble globular [GADV]-proteins exhibit what kind of catalytic activities?
(3) Can the first genetic code, GNC, be established by stereospecific complex formation between [GADV]-amino acid and GNC-containing oligonucleotide?
(4) Can single-stranded $(GNC)_n$ gene be produced by joining GNC in the complex?
(5) Can [GADV]-protein be produced by expression of single-stranded $(GNC)_n$ gene?
(6) Can double-stranded $(GNC)_n$ gene be produced by complementary strand synthesis of the single-stranded gene with [GADV]-protein?
(7) Can double-stranded $(GNC)_n$ gene be replicated with [GADV]-protein? and so on.

Points 6 and 7 can be described in other words as followings: Is there an RNA polymerase protein that is made from only GADV amino acids?

As described above, almost all evolutionary steps from [GADV]-protein world to the emergence of life remains experimentally unconfirmed, although that a hypothesis on the origin of life remains unconfirmed with experimental procedures does not mean that life did not emerge according to the hypothesis. However, GADV hypothesis is not purely theoretical idea, because GADV hypothesis was obtained on the basis of computer analyses of microbial gene and protein databases, in which experimental results were collected. Furthermore, evolutionary process to the emergence of life remains rationally unexplained with all ideas including RNA world hypothesis other than GADV hypothesis. Taking them compositely into consideration, I would like to conclude again that life emerged from [GADV]-protein world.

Acknowledgments: We are very grateful to Dr. Tadashi Oishi (G&L Kyosei Institute, Emeritus professor of Nara Women's University) for encouragement throughout our research on the origin and evolution of the fundamental life system.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gilbert, W. The RNA world, *Nature* **1986**, *319*, 618.
2. 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.
3. Ikehara, K. Possible steps to the emergence of life: The [GADV]-protein world hypothesis. *Chem. Rec*. **2005**, *5*, 107–118.
4. Imai, E.; Honda, H.; Hatori, K.; Brack, A.; Matsuno, K. Elongation of oligopeptides in a simulated submarine hydrothermal system. *Science* **1999**, *283*, 831–833.
5. Holm, N.G.; Andersson, E. Hydrothermal simulation experiments as a tool for studies of the origin of life on Earth and other terrestrial planets: A review. *Astrobiology* **2005**, *5*, 444–460.
6. Chandru, K.; Obayashi, Y.; Kaneko, T.; Kobayashi, K. Formation of amino acid condensates partly having peptide bonds in a simulated submarine hydrothermal environment. *Viva Origino* **2013**, *41*, 24–28.
7. Maury, C.P.; Origin of life. Primordial genetics: Information transfer in a pre-RNA world based on self-replicating beta-sheet amyloid conformers. *J. Theor. Biol.* **2015**, *382*, 292-297.
8. Sharov, A.A. Coenzyme world model of the origin of life. *Biosystems* **2016**, 144, 8-17.

9. De Farias, S.T.; Rego, T.G.; Jose, M.V. tRNA core hypothesis for transition from RNA world to the ribonucleoprotein world. *Life(Basel)* **2016**, *6*, pii: E15. Doi: 10.3390/life6020015.
10. Kruger, K.; Grabowski, P.J.; Zaug, A.J.; Sands, J.; Gottschling, D.E.; Cech, T.R. Self-splicing RNA: autoexcision and autocyclization of ribosomal RNA intervening sequence of *Tetrahymena*. *Cell* **1982**, *31*, 147-157.
11. Guerrier-Takada, C.; Gardiner, K.; Marsh, T.; Pace, N.; Altman, S. The RNA moiety of ribonuclease P is catalytic subunit of the enzyme, *Cell* **1983**, *35*, 849-857.
12. Ellington, A.D.; Szostak, J.W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **1990**, *346*, 818–822.
13. Tuerk, C.; Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **1990**, *249*, 505–510.
14. Thodima, V.; Pirooznia, M.; Deng, Y.; RiboaptDB: a comprehensive database of ribozymes and aptamers. *BMC Bioinformatics* **2006**, 7 Suppl 2:S6.
15. Bartel, D.P.; Szostak, J.W. Isolation of new ribozymes from a large pool of random sequences. *Science* **1993**, *261*, 1411-1418.
16. Ekland, E.H.; Szostak, J.W.; Bartel, D.P. Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science* **1995**, *269*, 364-370.
17. Watson, J.D.; Crick, F.H.C. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* **1953**, *171*, 737-738.
18. Dill, K.A. Dominant forces in protein folding. *Biochemistry* **1990**, *29*, 7133–7155.
19. Ikehara, K. Pseudo-replication of [GADV]-proteins and origin of life. *Int. J. Mol. Sci.* **2009**, *10*, 1525–1537.
20. 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.
21. Oba, T.; Fukushima, J.; Maruyama, M.; Iwamoto, R.; Ikehara, K. Catalytic activities of [GADV]-peptiders. Formation and establishment of [GADV]-protein world for the emergence of life. *Orig. Life Evol. Biosph.* **2005**, *35*, 447–460.
22. Jacobsen, N. E.; Pfaltz, A.; Yamamoto, H. eds. In *Comprehensive Asymmetric Catalysis* Springer, New York, U.S.A. 1999.
23. Shibasaki, M.; Yoshikawa, N.; Matsunaga, S. In *Comprehensive Asymmetric Catalysis*, eds. Jacobsen N.E., Pfaltz A. & Yamamoto H. Springer, New York, U.S.A. Suppl. 1, pp. 135‑142. 2004.
24. Gorlero, M.; Wieczorek, R.; Adamala, K.; Giorgi, A.; Schininà, M.E.; Stano, P.; Luisi, P.L. Ser-His catalyses the formation of peptides and PNAs. *FEBS Lett.* **2009**, *583*, 153–156.
25. 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. *Nucleic Acids Res.* **1996**, *24*, 4249–4255.
26. Ikehara, K.; Yoshida, S. SNS hypothesis on the origin of the genetic code; *Viva Origino* **1988**, *26*, 301-310.
27. Luisi, P.L. A new start from ground zero? *Orig. Life Evol. Biosph.* **2014**, *44*, 303–306.
28. Ikehara K. In *GADV Hypothesis on the Origin of Life -Life emerged in this way!?-*, LAMBERT Academic Publishing: Saarbrucken, Germany. 2016.
29. 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.
30. Ikehara, K. Protein ordered sequences are formed by random joining of amino acids in protein 0^{th}-order structure, followed by evolutionary process. *Orig. Life Evol. Biosph.* **2014**, *44*, 279–281.
31. Rode, B.M.; Peptides and the origin of life. *Peptides* **1999**, *20*, 773–786.

32. Plankensteiner, K.; Reiner, H.; Rode, B.M. Amino acids on the rampant primordial Earth: Electric discharges and the hot salty ocean. *Mol. Divers*. **2006**, *10*, 3–7.
33. Ruiz-Mirazo, K.; Briones,C.; de la Escosura, A. Prebiotic systems chemistry: new perspectives for the origins of life. *Chem. Rev*. **2014**. *114*, 285-366.
34. Chiarabelli, C.; Vrijbloed, J.W.; De Lucrezia, D.; Thomas, R.M.; Stano, P.; Polticelli, F.; Ottone, T.; Papa, E.; Luisi, P.L.; Investigation of *de novo* totally random biosequences, Part II: on the folding frequency in a totally random library of *de novo* proteins obtained by phage display. *Chem. Biodivers*. **2006**, *3*, 840-859.
35. Chessari, S.; Thomas, R.; Polticelli, F.; Luisi, P.L. The production of *de novo* folded proteins by a stepwise chain elongation: a model for prebiotic chemical evolution of macromolecular sequences. *Chem. Biodivers*. **2006**, *3*, 1202-1210.
36. Stryer, L. In *Biochemistry*, 3rd ed.; W.H. Freeman and Company: New York, NY, USA. 1988.
37. Guimaraes, R.C. Self-Referential Encoding on Modules of Anticodon Pairs-Roots of the Biological Flow System. *Life (Basel)* **2017**, *7*, doi: 10.3390/life7020016.
38. Miller, S.L.; Orgel, L.E. In *The Origins of Life on the Earth,* Prentice-Hall: Englewood Cliffs, NJ, USA, 1974.
39. Botta, O.; Glavin, D.P.; Kminek, G.; Bada, J.L. Relative amino acid concentrations as a signature for parent body processes of carbonaceous chondrites. *Orig. Life Evol. Biosph*. **2002**, *32*, 143–163.
40. Van der Gulik, P.T.; Speijer, D. How amino acids and peptides shaped the RNA world. *Life* **2015**, *5*, 230–246.
41. Shimizu, M. Specific aminoacylation of C4N hairpin RNAs with the cognate aminoacyl-adenylates in the presence of a dipeptide: origin of the genetic code. *J. Biochem*., (Tokyo). **1995**, *117*, 23-26.
42. Yoneda, S.; Shimizu, M.; Go, N.; Fujii, S.; Uchida, M.; Miura, K.; Watanabe, K. Theoretical and experimental approach to recognition of amino acid by tRNA. *Nucleic Acids Symp Ser*. **1983**, *12*, 145-148.
43. Prywes, N.; Blain, J.C.; Del Frate, F.; Szostak, J.W. Nonenzymatic copying of RNA templates containing all four letters is catalyzed by activated oligonucleotides. *eLife* **2016**, *5*, e17756. DOI: 10.7554/eLife.17756
44. Horning, D.P.; Joyce, G.F. Amplification of RNA by an RNA polymerase ribozyme. *Proc. Natl. Acad. Sci. USA*. **2016**, *113*, doi: 10.1073/pnas.1610103113.
45. Krupkin, M.; Matzov, D.; Tang, H.; Metz, M.; Kalaora, R.; Belousoff, M.J.; Zimmer man, E.; Bashan, A.; Yonath, A. A vestige of a prebiotic bonding machine is functioning within the contemporary ribosome. *Philos. Trans. R. Soc Lond. B. Biol. Sci*. **2011**, *366*, 2972-2978. doi: 10.1098/rstb.2011.0146.
46. Davidovich, C.; Belousoff, M.; Welselman, I.; Shapira, T.; Krupkin, M.; Zimmerman, E.; Bashan, A.; Yonath, A. The Proto-Ribosome: an ancient nano-machine for peptide bond formation. *Isr. J. Chem*. **2010,** *50*, 29-35
47. Sun, L.; Cui, Z.; Gottlieb, R.L.; Zhang, B. A selected ribozyme catalyzing diverse dipeptide synthesis. *Chem. Biol*. **2002**, *9*, 619-628.
48. Vicence, Q.; Cech, T.R. A natural ribozyme with 3',5' RNA ligase activity. *Nat. Chem. Biol*. **2009**, *5*, 97-99..
49. White, H.B. Coenzymes as fossils of an earlier metabolic state. *J. Mol. Evol*. **1975**, *7*, 101–104.
50. Cernak, P.; Sen, D. A thiamin-utilizing ribozyme decarboxylates a pyruvate-like substrate. *Nat. Chem*. **2013**, *5*, 971-977. doi: 10.1038/nchem.1777.