

Review

Not peer-reviewed version

How was the First Gene formed in the absence of GeneCitation: To be added by editorial staff during production.

[Kenji Ikehara](#) *

Posted Date: 25 July 2023

doi: 10.20944/preprints202307.1624.v1

Keywords: origin of gene; anticodon stem-loop tRNA (AntiC-SL tRNA); immature [GADV]-protein; protein 0th-order structure; origin of life; GADV hypothesis



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Review

How Was the First Gene Formed in the Absence of Gene

Kenji Ikehara ^{1,2,*}

¹ G&L Kyosei Institute, In the Keihanna Academy of Science and Culture (KASC), Keihanna interaction plaza, Lab. Wing 3F, 1-7 Hikaridai, Seika-cho, Souraku, Kyoto 619-0237, Japan;

² International Institute for Advanced Studies, Kizugawadai 9-3, Kizugawa, Kyoto 619-0225, Japan

* Correspondence: ikehara@cc.nara-wu.ac.jp; Tel.: +81-774-73-4478

Simple Summary: In GADV hypothesis, it is considered that life emerged after six members composing of the fundamental life system were formed in order of protein, cell structure, metabolism, tRNA, genetic code and gene. This means that the first gene must be formed in the absence of gene or under the five members except gene. There is apparently contradiction between formation of the first gene and the five members used for generation of the first gene, because the five members have been produced under genetic function in extant organisms. However, it can be reasonably explained by considering that gene was generated by five prototypes of members, as immature [GADV]-protein, [GADV]-microsphere, metabolism using immature [GADV]-proteins, Anticodon stem-loop (AntiC-SL) tRNA, GNC primeval genetic code.

Abstract: As a matter of course, the first gene must be formed in the absence of gene. On the other hand, many biopolymers including gene are produced under the genetic system in extant organisms. Thus far, no idea explaining how the first gene was formed in the absence of gene, has been proposed except the idea based on the GADV hypothesis. GADV means four amino acids; Gly [G], Ala [A], Asp [D] and Val [V]. In this article, a reliable answer to the question, how the first gene was generated, is provided. The idea is as follows. The first gene (genetic information) was formed by random joining of anticodons, GNCs, which were carried by primeval anticodon stem-loop (AntiC-SL) tRNAs produced during repeated cycles of random joining of nucleotides and degradation of oligonucleotides. However, it might be difficult for many persons to accept the idea, because tRNAs and metabolic pathways using various proteins, which are produced under genetic functions, must be used to form the first gene. Then, it is reconsidered based on definition of genetic information and minimum necessary things to generate the first gene in this article, how the first gene was generated. Consequently, it could be reconfirmed that five prototypes of members (genetic code, tRNA, metabolism, cell structure and protein) composing the fundamental life system, except gene, are certainly necessary to generate the first gene. Namely, it can be concluded that the first gene was formed through nucleotide metabolism with immature [GADV]-proteins, primeval AntiC-SL tRNA formation with immature [GADV]-proteins, formations of single-stranded (GNC)_n codon sequence by random joining of anticodons, GNCs, carried by AntiC-SL tRNAs, double-stranded (ds)-(GNC)_n RNA and maturation of an immature [GADV]-protein, which was produced from one strand of the ds-RNA. Thus, the first gene was formed using the five primitive or prototypes of members, which were produced in the absence of gene. There would be no other way for generating the first gene.

Keywords: origin of gene; anticodon stem-loop tRNA (AntiC-SL tRNA); immature [GADV]-protein; protein 0th-order structure; origin of life; GADV hypothesis;

1. Introduction

Many organisms are living using various types of beautiful mature proteins on the present Earth. However, it is unknown how the first gene or the first genetic information, which should trigger to

produce modern genes for synthesizing such beautiful proteins, was generated, although elucidation of the mechanism must be one of the most important problems in molecular biology and biochemistry.

In this article, it is discussed based on [GADV]-protein world hypothesis or GADV hypothesis [1–3] how the first (GNC)_n gene encoding a mature [GADV]-protein was generated in the absence of gene.

On the other hand, the following ideas regarding the first gene or the smallest gene have been proposed.

1. H. S. Bernhardt *et al.* have presented the idea that Poly-G is the most primitive polynucleotide, which is encoded to poly-Gly by being translated three bases at a time [4–6].
2. Another hypothesis on the first gene has been proposed by C. J. Michel and J. D. Thompson that X circular code represents a possible ancestor of the standard genetic code. It is considered under the hypothesis that the first gene was repeating sequence of the following three codons, {GGT;GTA;GTC}, since words of the code appear only in the reading frame [7].
3. The idea has been also proposed that polynucleotide encoding Microcin C7, a modified linear heptapeptide, Acetyl-Met-Arg-Thr-Gly-Asn-Ala-Asp-X, where X symbolizes an unidentified, acid-labile group substituting for the C-terminal aspartate residue, is the smallest gene [8,9].

Hence, definitions of some terms, such as gene and protein, which are appeared in this article, are first given in Table 1 for readers of this article to avoid confusions and to make it easier to understand the discussion about how the first gene was generated.

Some ideas as the first gene or the smallest gene, which are different from those described in this article, have been proposed thus far. Therefore, definitions of GADV, gene, protein, codon etc. are described in this Table for readers to make it easier to understand discussion described in this article without avoid confusions.

Table 1. Definitions of gene, protein and codon.

Term	Definition
GADV	GADV means four amino acids; Gly [G], Ala [A], Asp [D] and Val [V], which are manifested in one-letter symbols.
Gene	"Gene" means a region of RNA or DNA encoding the smallest matured water-soluble globular protein composed of around 100 amino acid residues.
Mature gene	"Mature gene" is a gene encoding a mature water-soluble globular protein composed of around 100 amino acid residues.
Immature gene	"Immature gene" is a gene encoding an immature and flexible water-soluble globular protein composed of around 100 amino acid residues.
Protein	"Protein" means a water-soluble globular protein composed of around 100 amino acid residues.
Mature protein	"Mature protein" is a protein like a precision polymer machine, which was formed by maturation from an immature protein.
Immature protein	"Immature protein" is a protein, which was produced by random joining of [GADV]-amino acids in protein 0 th -order structure or from antisense strand of a GC-rich gene, as (GNC) _n , (SNS) _n and a modern GC-rich gene.
Codon	"Codon" means three bases (triplet) encoding an amino acid, but not singlet or doublet.

1.1. Ideas on the origin of gene, which have been considered so far

Elucidation of the origin of gene is one of important key steps leading to solving the “mystery” of the origin of life, because gene is one of the six members (gene, genetic code, tRNA, metabolism, cell structure and protein) of the fundamental life system [1]. In fact, two main ideas on the origin of gene

have been proposed thus far. Those are “gene duplication theory”, which was proposed by S. Ohno [10] and “exon-shuffling theory” presented by W. Gilbert [11]. However, genes previously existed are required to form new genes in both theories. Therefore, it must be stated that those are not genuine theories for explaining the origin of gene, because the formation mechanism of the first gene could not be explained by the two theories.

1.2. Ideas on the origin of gene considered in RNA world hypothesis

The “mystery” of the origin of life has also been unsolved still now, irrespective of strenuous efforts of many researchers. It has become one of the large obstacles for solving the mystery that the mechanism, how the first gene for protein synthesis was formed during random reactions on the primitive Earth, could not be clarified [1]. For example, in RNA world hypothesis [12], which is classified in “gene/replicator-early theory”, it has been only vaguely considered that genetic information should be formed one day in self-replicated RNAs, which are mere genetic information carriers. However, it would be impossible to write genetic information into the self-replicated RNA, even if RNA could be self-replicated, because it must be principally impossible to write genetic information of a mature protein into such self-replicated RNA in the absence of protein, and also because, of course, it is obvious not to be able to design an amino acid sequence of a mature protein in advance. In fact, not the origin of gene but only self-replication of RNA has been discussed in the studies on RNA world hypothesis thus far [13,14].

1.3. The reason why the origin of gene has been unsolved still now

Then, how was the first genetic information written in RNA? In this article, it is described as assuming that the first genetic information carrier, which was used when life emerged, was RNA but not DNA as generally considered. It is well known that both a mature gene and a mature protein never be produced by random joining of the respective monomers because of the extraordinary large sequence diversities, $(4^3)^{100} = \sim 10^{180}$ and $20^{100} = \sim 10^{130}$, respectively [1,15]. This means that any gene encoding a mature protein never be generated at one stroke through a random process. Furthermore, there is another reason. That is because nobody has noticed that it is impossible to form the first gene or an entirely new gene without an immature protein necessary to form a mature protein [1]. In other words, it would be principally impossible to write genetic information in RNA independently of a codon sequence for synthesis of an immature protein, which is produced by random joining of [GADV]-amino acids in [GADV]-amino acid composition or under a protein 0th-order structure [1,16].

In this article I would like to aim at exactly discussing formation mechanism of the first gene coding for an amino acid sequence of a mature protein. Therefore, first the origin of gene is discussed more strictly than before based on definition of genetic information and necessities for generating the first gene in order to solve the genuine origin of gene. Then, the idea about the origin of life in GADV hypothesis [1–3], which I have proposed, including formation process of the first gene is described in a nutshell in the next Session.

2. Steps to the Emergence of Life considered in GADV hypothesis

As shown in Figure 1, it is considered that life emerged as piling up six members one by one along the time axis from chemical evolution via formation process of the first gene [1].

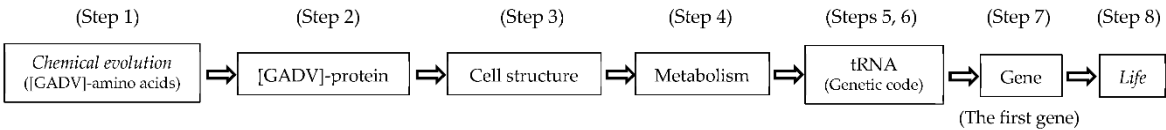


Figure 1. Formation process of the first gene (genetic information), which led to the emergence of life. Note that the process began from immature [GADV]-protein formed by random joining of [GADV]-amino acids. The first (*Chemical evolution*) and the last (*Life*) steps in the eight steps are described in *Italic letters*.

The eight steps from chemical evolution to the emergence of life are overviewed below, because the detailed steps were already described in the book published from Springer Nature [1] (Figure 1, Table 2).

Step 1. Chemical evolution

[GADV]-amino acids were synthesized with various prebiotic means and accumulated on the primitive Earth.

Step 2. Production of immature [GADV]-proteins as [GADV]-peptide aggregates

Immature [GADV]-proteins as [GADV]-peptide aggregates were produced by direct random joining of [GADV]-amino acids in protein 0th-order structure (a special amino acid composition containing [GADV]-amino acids at roughly equal amounts) during repeated wet-dry cycles.

Step 3. Formation of [GADV]-microsphere

[GADV]-microspheres were formed with immature [GADV]-proteins [17]. The formation of [GADV]-microspheres made it possible to select and evolve them owing to acquisition of individuality of microspheres.

Step 4. Formation of proto-metabolism

[GADV]-peptides could be synthesized by immature [GADV]-proteins using [GADV]-amino acids incorporated from circumstances into [GADV]-microspheres. Furthermore, syntheses of nucleobases and ribose-5-phosphate made it possible to produce nucleotides through proto-metabolic system using immature [GADV]-proteins in [GADV]-microspheres [1]. Accompanied by the syntheses, oligonucleotides were also produced by random joining of nucleotides.

Step 5. Formation of anticodon-stem loop (AntiC-SL) tRNA

AntiC-SL-tRNAs were produced by repeated random joining of nucleotides and degradation of oligonucleotides synthesized through catalytic activities of immature [GADV]-proteins [1]. A gene for synthesis of an AntiC-SL tRNA was formed at this time point. Immature [GADV]-proteins could be produced more efficiently by using AntiC-SL-tRNAs than before. Thus, synthetic system of immature proteins evolved in a bid to produce immature [GADV]-proteins more efficiently in [GADV]-microspheres, which equipped with only an incomplete genetic system. Actually, the synthetic system could evolve step by step because [GADV]-microspheres having more efficient system for synthesis of immature [GADV]-proteins could produce more descendants than others.

Step 6. Establishment of GNC primeval genetic code

GNC primeval genetic code was established using [GADV]-amino acids and GNC codons as assumed by GNC code frozen-accident theory [1,18].

Step 7. Formation of the first gene (genetic information) encoding an ordered [GADV]-amino acid sequence for formation of a mature [GADV]-protein

It is here explained about formation of the first gene in some detail, because discussion about the origin of gene or generation of the first gene is the purpose of this article.

A single-stranded (ss)-(GNC)_n RNA encoding a random [GADV]-amino acid sequence of an immature [GADV]-protein was formed by random joining of four anticodons, GNCs, which were carried by four nonspecific AntiC-SL tRNAs [1]. Successively, a double-stranded (ds)-RNA was formed by complementary strand synthesis of the ss-(GNC)_n RNA. The formation of ds-(GNC)_n RNA triggered to generate the first gene for synthesis of a mature [GADV]-protein. Genetic information of the first gene for a mature [GADV]-protein was formed through maturation of the immature [GADV]-protein, which was produced by expression of a random (GNC)_n sequence on one strand of the ds-(GNC)_n RNA [1].

Step 8. The emergence of life

It is considered that the first life with a genetic system emerged when necessary number of genes were prepared [1]. Note that it would be impossible to determine exactly the moment of the emergence of life, because various activities in [GADV]-microsphere should be still expressed by many immature [GADV]-proteins, when the first gene was formed.

It must be recognized that the steps never proceed inversely. One of the reasons, why origins of gene and life could not be solved thus far, would be because nobody did comprehend correctly the hierarchy among formations of immature protein, AntiC-SL tRNAs and gene.

I am naturally confident of GADV hypothesis, assuming that the first gene was formed through Step 1: Chemical evolution to Step 7: Formation of the first gene (Figure 1, Table 2), because the steps from chemical evolution to the emergence of life can be reasonably explained [1]. However, it might be difficult for many readers to accept the idea, because the generation process of the first gene includes many steps, which it seems almost impossible to realize the processes such as establishment of metabolic system for synthesis of nucleotides, formations of AntiC-SL tRNAs and ds-(GNC)_n RNA and so on. The reason might be also probably because many members, which are produced under genetic system including gene in extant organisms, are used to form the first gene in the absence of gene. In other words, the so-called complete genetic system was absent until the first gene was formed. Then, in the following Sections, it is confirmed whether or not those members must be really required to form the first gene. For the purpose, it is described what were indispensable to form the first gene in more detail as focusing on the formation process of the first gene. Needless to state, the first gene, which is discussed in this article, is a gene encoding an amino acid sequence of a mature [GADV]-protein (Table 1), because genes for synthesis of AntiC-SL tRNAs had been already formed before the first gene for protein synthesis was generated (see Step 5 in Section 2 above).

Table 2. Evolutionary process to Step 8: emergence of life from Step1: Chemical evolution. The first gene was formed at the Step 7: Gene. “imm-” means “immature” in this Table.

Step	Materials synthesized, etc.	Phenomena occurred at the step
Step 1: <i>Chemical evolution</i>	[GADV]-amino acids	[GADV]-amino acids in messy environments [19]
Step 2: Protein	imm-[GADV]-proteins	Direct random joining of [GADV]-amino acids (in protein 0th-order structure [1,16])
Step 3: Cell structure	[GADV]-microsphere	[GADV]-protein membrane ([GADV]-protein world) Catalytic reactions with imm-[GADV]-proteins
Step 4: Metabolism	Proto-metabolism	Synthesis of [GADV]-peptides (imm-[GADV]-proteins) For the synthesis, aa-AMP and aa-3'-ACC-5' were used as activated [GADV]-amino acids in order. Synthesis of nucleotides
Step 5: tRNA	AntiC-SL tRNA	Use of proto-tRNAs (AntiC-SL) 1. Dimer of AntiC-SL tRNAs (juxtaposed) 2. Tetramer of AntiC-SL tRNAs
Step 6: Genetic code	GNC primeval genetic code	(GNC code frozen accident theory)
Step 7: Gene	ds-(GNC) _n gene	1. Single-stranded (GNC) _n RNA 2. Double-stranded (GNC) _n RNA 3. Double-stranded (GNC) _n RNA (genetic information)
Step 8: Life	The emergence of life	Note: Life did not always arise upon formation of only one gene.

3. The reason why it is difficult to accept the formation process of the first gene

I am naturally confident of GADV hypothesis, assuming that the first gene was formed through Step 1: Chemical evolution to Step 7: Formation of the first gene (Figure 1, Table 1), because the steps from chemical evolution to the emergence of life can be reasonably explained [1]. However, it might be difficult for many readers to accept the idea, because the generation process of the first gene includes many steps, which it seems almost impossible to realize the processes such as establishment of metabolic system for synthesis of nucleotides, formations of AntiC-SL tRNAs and ds-(GNC)_n RNA and so on. The reason might be also probably because many members, which are produced under genetic system including gene in extant organisms, are used to form the first gene in the absence of gene. In other words, the so-called complete genetic system was absent until the first gene was formed. Then, in the following Sections, it is confirmed whether or not those members must be really required to form the first gene. For the purpose, it is described what were indispensable to form the first gene in more detail as focusing on the formation process of the first gene. Needless to state, the first gene, which is discussed in this article, is a gene encoding an amino acid sequence of a mature [GADV]-protein, because genes for synthesis of AntiC-SL tRNAs had been already formed before the first gene for protein synthesis was generated (see Step 5 in Section 2 above).

4. Definition of Gene used in Modern Organisms

The more precise definition of gene or genetic information than that given in Table 1 as a prerequisite for discussing the formation process of the first gene is described below in order to reconsider going back to the basis that the steps shown in Figure 1 and Table 2 were really indispensable to form the first gene. Gene is defined as follows.

- 1. A region on RNA or DNA, where codon sequence for synthesis of an amino acid sequence of a mature protein is written.
- 2. Gene for synthesis of a mature protein is written as a commaless triplet codon sequence in ds-RNA or DNA under a genetic code (Figure 2).

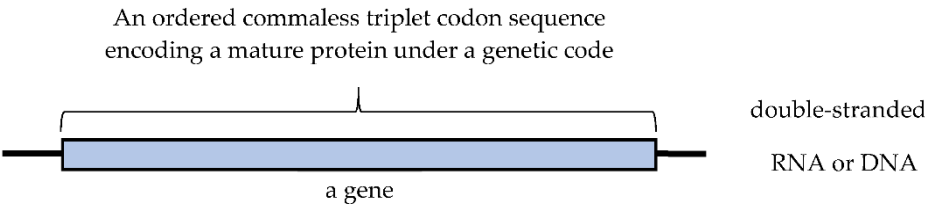


Figure 2. Genetic information is defined by a commaless codon sequence for synthesis of a mature protein, which is written under a genetic code in ds-RNA or DNA as a genetic information carrier. The main problem in this article is how the first genetic information was generated.

Note here that the words, commaless triplet “codon sequence” and a “mature protein”, but not “base sequence” and a “protein” are used respectively in the above sentence. The reasons are because gene is in fact written as a “codon” sequence but not a “base” sequence. Furthermore, the reason why the phrase, a “mature” protein, is explicitly used in the definition, is because a mature protein or a general protein, which is encoded by a gene, must be distinguished from an “immature” protein, which is produced through a type of random process, for example, by direct random joining of [GADV]-amino acids in protein 0th-order structure [1,16]. In addition, it is necessary to introduce an immature protein in order to write a genetic information of a mature protein into ds-RNA, because maturation process of an immature protein is essential in order to acquire genetic information of a mature protein (Figure 3 (B)). Therefore, codon sequence on a ss-RNA corresponding to a modern mRNA is eliminated from the definition of gene, because mRNA is mere genetic information carrier, which is only transcribed from ds-RNA or DNA and the codon sequence of mRNA can not generally used to write genetic information into ds-RNA or DNA except exceptional retro-transcription.

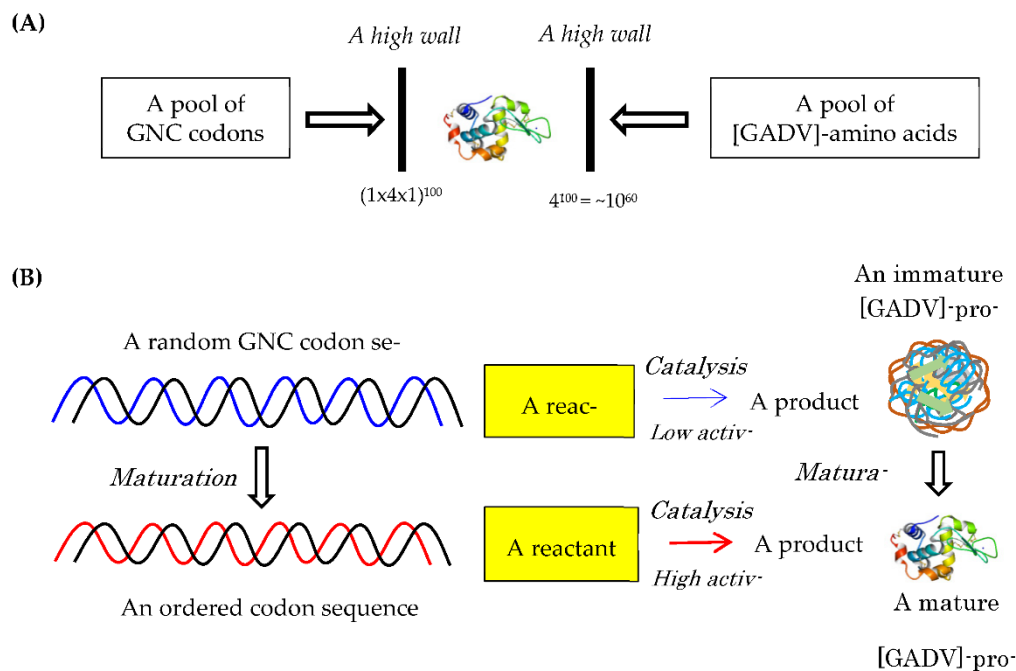


Figure 3. (A) Both a mature protein and a mature gene encoding a mature protein never be produced by random joining of [GADV]-amino acids and GNC codons encoding [GADV]-amino acids respectively in the GNC code era, because of extremely high walls as $\sim 10^{60} = 4^{100}$. **(B)** The process for acquiring genetic information of an entirely new mature protein. The entirely new genetic information including the first gene must be always generated through maturation process from an immature protein, which is produced by expression of an essentially random codon sequence, to a mature protein. Furthermore, the process must be led by elevation of a weak catalytic activity on the immature protein to a mature protein with a high activity. Blue and red wavy lines show RNA strand encoding a random (GNC)_n codon sequence for synthesis of an immature protein and an ordered (GNC)_n codon sequence encoding a mature protein, respectively.

As a matter of course, it must be also paid attention to that, even if gene composed of a codon sequence could be formed, the codon sequence is meaningless if genetic code for translation of the sequence is absent (although, of course, the gene never be formed in the absence of genetic code), because the codon sequence cannot be translated into a protein in the absence of genetic code.

5. Necessary conditions for generating the first gene

Needless to state, both a mature gene and a mature protein never be produced by random joining of the respective monomers, codons and amino acids, because there are extraordinary high walls of 10^{60} even in the GNC code era [1], which it is substantially impossible to pass over (Figure 3 (A)). Then, consider the conditions necessary for generating the first gene, as referring the above definition of gene (Figure 2), because the formation process, how the first gene was created in the absence of gene, could be made clear using the conditions (Figure 3 (B)).

1. It is indispensable to start from random commaless GNC codon sequence under GNC primeval genetic code to form the first gene because a [GADV]-amino acid sequence for a mature [GADV]-protein never be designed in advance.
2. It is necessary to produce an immature [GADV]-protein using the random commaless codon sequence as the start point for evolving it to a mature [GADV]-protein. Note that the random codon sequence codes for an immature but meaningful water-soluble globular [GADV]-protein with a random [GADV]-amino acid sequence arranged under a protein 0th-order structure or from [GADV]-amino acid composition.
3. Minimal proto-type of translation system (AntiC-SL tRNA and GNC primeval genetic code) is also indispensable to express the codon sequence encoding the random [GADV]-amino acid

sequence in order to produce an immature [GADV]-protein according to the random GNC codon sequence.

4. Furthermore, it is necessary to use a codon sequence on a ds-RNA for memorizing base changes appeared during the evolutionary process.
5. Cell structure is also necessary to select a codon sequence encoding an immature [GADV]-protein with a higher function than before, which was produced during the evolutionary process.

From the above conditions for generating the first gene, it would be understandable that the Steps from 2 to 6 drawn in Figure 1 and Table 2 which it might be intuitively difficult to accept, are indispensable to generate the first gene. Thus, not a complete but an immature (proto-type) translation system is undoubtedly required to generate the first gene.

Therefore, the first gene for a mature [GADV]-protein must be always formed through maturation of an immature water-soluble globular [GADV]-protein with some flexibility as adjusting an incomplete active site on the immature [GADV]-protein to fit the substrate (Figure 3 (B)). This could be easily understood from an example that a key and a keyhole, which must be firmly matched each other, cannot be formed by construction of a key and a keyhole both of which are randomly made without referring the respective structures. In fact, a key should be constructed as fitting a mold of the corresponding keyhole. Similarly, a mature [GADV]-protein must be always produced from the corresponding immature [GADV]-protein through a type of evolutionary process from an immature [GADV]-protein with a weak active site to a mature [GADV]-protein having a site firmly fitting with the substrate. Thus, it could be understood that this is the only way for generating the first gene for formation of a mature [GADV]-protein (Figure 3 (B)). Inversely stating this, it is impossible to generate the first gene without the Steps from 2 to 6. This means that any gene/replicator-early theory like as RNA world hypothesis [12], which does not predict an evolutionary pathway from an immature to a mature protein, is not a valid idea explaining the origin of life.

Then, consider concretely the formation process of the first gene (Figure 4). A ss-(GNC)_n RNA must be produced by random joining of GNC anticodons carried by AntiC-SL tRNAs. Next, an immature [GADV]-protein produced from a (GNC)_n codon sequence on one strand of ds-(GNC)_n RNA, which was formed by complementary strand synthesis of the ss-(GNC)_n RNA, must be matured to a [GADV]-protein like a precision polymer machine (Figure 3 (B)).

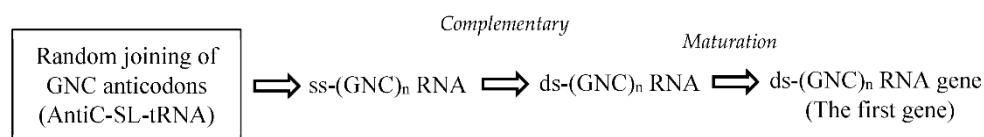


Figure 4. Formation process of the first ds-(GNC)_n RNA gene (Anticodon joining hypothesis [1]).

6. Three keys for opening doors leading to formation of the first gene

It should be noted that a random [GADV]-amino acid sequence encoded by the random (GNC)_n codon sequence is the substantially same with a random [GADV]-amino acid sequence, which was produced by direct random joining of [GADV]-amino acids in protein 0th-order structure. Therefore, it is expected that a water-soluble globular immature [GADV]-protein with some flexibility could be formed by expression of the random (GNC)_n codon sequence at a high probability. Therefore, ds-RNA, AntiC-SL tRNAs, nucleotide metabolic system and so on were certainly necessary to form the first gene, although of course those were prototype of members, which are quite different from those usually produced under the modern genetic system including gene. On the other hand, those things as can be seen in the steps from chemical evolution to the formation of the first gene could be surely produced without involvement of genes (Figure 1). Therefore, it means that the problems remain unsolved how ds-RNA, AntiC-SL tRNAs and nucleotide metabolic system could be produced without involvement of genes.

The difficult problems could be solved by the three keys, all of which must be produced through random processes in the absence of gene, however, are indispensable things to generate the first gene. The three keys are (1) Immature [GADV]-protein produced by random joining of [GADV]-amino acids under protein 0th-order structure. (2) AntiC-SL tRNA, which was formed through cycles of random joining of nucleotides and degradation of the oligonucleotides. (3) ss-(GNC)_n RNA, which was produced by random joining of anticodons, GNCs, carried by AntiC-SL tRNAs. The first gene could be generated owing to the three keys.

It might be still difficult for many readers to accept the scenario given in Figure 1. However, life did emerge on the primitive Earth about 4 billion years ago and versatile organisms have flourished on the present Earth. The facts clearly indicate that the first gene was surely formed one day about 4 billion years ago. Therefore, if no process leading to formation of the first gene exists other than the process drawn in Figure 1, we cannot help but consider that the first gene was generated along the formation process drawn in Figure 1 and that the formation process, which is considered based on the definition of gene (Figure 2), is the indispensable one necessary for forming the first gene (Figure 3 (B)), must be the correct answer to the problem how the first gene was formed.

7. Discussion

As a matter of course, no gene existed before generation of the first gene. Therefore, any complete genetic function could not be asked at the time until the first gene was generated. This means that the most primitive type of translation system (tRNAs and genetic code), metabolism and proteins, which were necessary to generate the first gene, must be produced through essentially random processes in the absence of gene. That is just the things or the six steps from Step 1: chemical evolution to Step 6: genetic code, except two steps, Step 7: gene and Step 8: life, which are drawn in Figure 1. The first gene was actually generated and life surely emerged on the primitive Earth, although it may not be easy to understand progression process of the steps. The three keys, 1. immature [GADV]-protein (protein 0th-order structure), 2. AntiC-SL tRNA and 3. ss-(GNC)_n RNA, made it possible to solve the difficult problems. It may be still difficult to accept the steps to the emergence of life. That may be also natural, because, otherwise, someone should have noticed the steps to the emergence of life before discovery of the process drawn in Figure 1. However, the fundamental life system including the core life system, which are composed of gene, tRNA, genetic code and protein, must be established by going against the stream of gene expression or from immature protein to the first gene via tRNA (genetic code), because formation of a system producing a complete object must always start from an incomplete object and should proceed towards forming more complete objects step by step. That is,

1. Genetic information (gene) encoding a mature protein cannot be formed in the absence of the object or an immature protein.
2. A mature object like a mature protein cannot be directly produced in the absence of genetic information (gene) (Figure 3 (A)).
3. The only way for solving the difficult problems is first to produce an immature protein with some flexibility and successively to form a mature protein alongside maturation of genetic information (Figure 3 (B)).

In this article, the formation process of the first gene has been logically explored according to the conditions for forming the first gene. Consequently, it could be confirmed that the formation of the first gene should progress from formation of immature [GADV]-proteins, formation of prototype of metabolism in [GADV]-microsphere, formation of AntiC-SL tRNAs, establishment of GNC primeval genetic code, finally to the formation of the first gene, as assumed by GADV hypothesis [1].

Next, consider formation process of new homologous genes. The formation process of new homologous genes has been well established as the duplication theory, which was proposed by S. Ohno [10]. The hypothesis assumes that new homologous genes encoding homologous proteins to an original protein are formed by accumulation of mutations onto the sense strand, which are

necessary to form a new homologous protein encoded by a new gene, after gene-duplication (Figure 5 (route 1)).

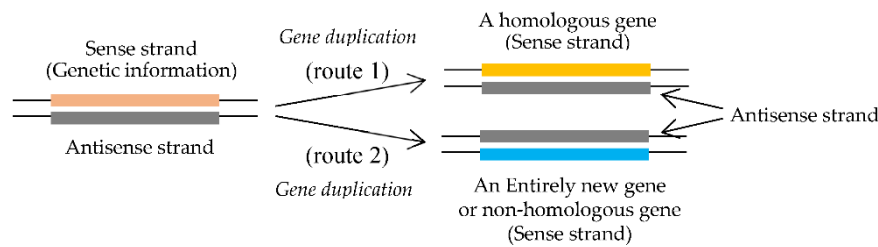


Figure 5. Formation processes of a new homologous gene from sense sequence of a gene (route 1) and an entirely new gene from antisense sequence of a GC-rich gene (route 2). Note that the formation process of an entirely new gene (route 2) always relies on maturation of an immature protein unlike route 1. For the formation of entirely new genes, three types of nonstop frames on antisense strands of GC-rich genes (GC-NSF(a)s or (GNC)_n-NSF(a), (SNS)_n-NSF(a) and GC-NSF(a), were used at every genetic code eras or GNC code, SNS code and universal genetic code. The respective protein 0th-order structures make it possible to form entirely new genes.

It must be taken notice that there exist two aspects of the origin of entirely new genes encoding an entirely new protein.

- (1) Formation process of the first gene, which led to the emergence of life: The one aspect is to clarify how the first gene or the first genetic information was formed, and, therefore, how the first gene was generated in the absence of any gene, as discussed in this article.
- (2) Formation process of entirely new genes after generation of the first gene: The other aspect is to make clear how entirely new genes have been created after formation of the first gene until the present day.

Entirely new genes except the first gene must be also formed through maturation process of an immature gene to a mature gene, as expected by pan-GC-NSF(a) hypothesis [1]. That is, entirely new genes were formed using codon sequences on antisense strands of 1. GC-rich (GNC)_n genes in GNC code era, 2. GC-rich (SNS)_n genes in SNS code era and 3. GC-rich genes in universal genetic code era [1,20]. The formation process of entirely new genes in the universal genetic code era was proposed as GC-NSF(a) hypothesis about 25 years ago [20]. The GC-NSF(a) hypothesis explains the generation process of new genes encoding an entirely new protein, which does not show meaningful sequence homology to any other previously existed proteins (Figure 5 (route 2)). Note that not only the first gene, which led to the emergence of life, but also entirely new genes, which are generated even in modern organisms, have been formed under the essentially same mechanism as shown in Figure 3 (B).

The reason, why many researchers had overlooked the second pathway for generation of entirely new genes, would be probably because AT-rich stop codons appear at a high frequency on an antisense sequence of genes having a GC content from about 30% to about 55%. Therefore, antisense sequences of AT-rich genes are generally inappropriate to code for an entirely new protein. That would be one of the reasons why the coding ability of an immature protein on antisense codon sequence of even a GC-rich gene (GC-NSF(a)) has been underestimated.

Formation of a homologous gene is independent of GC content of a parental gene but generation of an entirely new gene from an antisense sequence is restricted in GC-rich gene. The reason, why the field for generation of entirely new genes is restricted in GC-rich genes, is because one of protein 0th-order structures or one of codon sequences similar to GC-rich (SNS)_n-encoding 10 amino acids is written even in the GC-NSF(a)s of modern genes [1,20].

By considering pan-GC-NSF(a) hypothesis together with the gene-duplication theory, which opened the way for formation of homologous genes, it became possible to explain formation processes of all types of new genes or new homologous genes [10] and new non-homologous genes

(Figure 5) [1,20]. Thus, the two ways for formation of new homologous genes and of entirely new genes have made it possible to produce many new homologous genes and various entirely new genes. The reason why the pan-GC-NSF(a) hypothesis is also considered as a kind of duplication theory is because the process also requires duplication of the original genetic region to generate an entirely new protein (Figure 5 (route2)) similarly to the gene-duplication theory (Figure 5 (route 1)). Thus, genes encoding all types of new proteins necessary to live could be generated from sense sequences of all genes and antisense sequences of GC-rich genes (pan-GC-NSF(a), respectively (Figure 5), after formation of the first gene.

The ds-DNA structure, which is almost distortion-free, made it possible to provide infinite field for preserving a large number of two types of new genes by joining of ds-DNA in a column. That is also supported by extraordinary large amino acid sequence diversity of protein ($\sim 10^{130}$ in the universal genetic code era) and extraordinary base sequence diversity of DNA ($\sim 10^{180}$ in the universal genetic code era). The extraordinary sequence diversities, which have made it difficult to solve the mystery of the origins of gene and life, inversely make it possible to produce various proteins and genes and for versatile organisms to flourish on the present Earth.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

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

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

References

1. Ikehara, K. *Towards Revealing the Origin of life.—Presenting the GADV Hypothesis*; Springer Nature, Gewerbestrasse: Cham, Switzerland, **2021**.
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. Bernhardt, H.S.; Patrick, W.M. Genetic code evolution started with the incorporation of glycine, followed by other small hydrophilic amino acids. *J. Mol. Evol.* **2014**, *78*, 307–309.
5. Bernhardt, H.S.; Tate, W.P. Evidence from glycine transfer RNA of a frozen accident at the dawn of the genetic code. *Biol. Direct* **2008**, *3*, 53.
6. Inouyea, M.; Takino, R.; Ishida, Y.; Inouyea, K. Evolution of the genetic code; Evidence from serine codon use disparity in *Escherichia coli*. *Proc. Natl. Acad. Sci.* **2020**, *117*, 28572–28575.
7. Michel, C.J.; Thompson, J.D. Identification of a circular code periodicity in the bacterial ribosome: origin of codon periodicity in genes? *RNA Biol.* **2020**, *17*, 571–583.
8. García-Bustos, J.F.; Pezzi, N.; Mendez, E. *Antimicrob. Agents Chemother.* **1985**, *27*, 791–797.
9. González-Pastor, J.E.; José L. San Millán, J.L.; Moreno, F. The smallest known gene. *Nature*, **1994**, *369*, 281.
10. Ohno S. *Evolution by gene duplication*, Springer, Heidelberg, Germany, **1970**.
11. Gilbert W. The exon theory of genes. *Cold Spring Harb Symp Quant Biol* **1987**, *52*, 901–905.
12. Gilbert W. Origin of life: The RNA world. *Nature* **1986**, *319*, 618.
13. Joyce, G.F.; Szostak, J.W. Protocells and RNA Self-Replication. *Cold Spring Harb. Perspect. Biol.* **2018**, *10*, a034801.
14. Salditt, A.; Karr, L.; Salibi, E.; Le Vay, K.; Braun, D.; Mutschler, H. Ribozyme-mediated RNA synthesis and replication in a model Hadean microenvironment. *Nat Commun.* **2023**, *14*, 1495.
15. Dill, K.A. Dominant forces in protein folding. *Biochemistry* **1990**, *29*, 7133–7155.
16. 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.
17. Fox, S.W.; Dose, K. *Molecular evolution and the origin of life*, Marcel Dekker Inc., New York, USA.
18. Ikehara, K. Why Were [GADV]-amino Acids and GNC Codons Selected and How Was GNC Primeval Genetic Code Established? *Genes* **2023**, *14*, 375.

19. Ikehara, K. How Did Life Emerge in Chemically Complex Messy Environments? *Life* **2022**, 12:1319.
20. 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.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.