 doi:10.20944/preprints201912.0021.v1

Review

Protein 0th-Order Structure: The Key for Creating Entirely New Gene/Protein

Kenji Ikehara[1,2,3,*]

[1] *G&L Kyosei Institute, Koharu Bld. 202, Hokkeji 153-4, Nara 630-8001, Japan*

[2] *The International Institute for Advanced Studies of Japan, Kizugawadai 9-3, Kizugawa, Kyoto 619-0225, Japan*

[3] *Professor emeritus of Nara Women's University, Japan*

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

Abstract: Understanding the mechanism, how entirely new (EntNew) gene/protein or the first ancestral gene/protein of a family was created, should be one of the most important issues in the biological sciences. However, the mechanism is totally unknown still now. On the other hand, it is well known that mature protein is generally rigid and one catalytic center exists on the protein. Creation of such a mature EntNew gene/protein should be, of course, carried out through random process, because it cannot be designed in advance. However, the EntNew gene/protein never be created by random polymerization of the respective monomeric units, because of the extraordinary large sequence diversities of ~10^{180} and ~10^{130}, respectively. Protein 0th-order structure or a specific amino acid composition, in which immature but water-soluble protein can be produced even through random process, holds the key for solving the difficult problem. As it was fragmentally described in the previous papers how and where EntNew gene/protein was created, I describe in detail in this review three processes generating EntNew gene/protein with some flexibility under three genetic codes, the universal genetic code, SNS primitive code and GNC primeval code, and discuss why the mature gene/protein could be created through the processes.

Keywords: protein 0th-order structure; origin of gene; origin of protein; origin of genetic code; GNC primeval genetic code hypothesis; SNS primitive genetic code hypothesis

1. Introduction

In every textbook of "Biochemistry" or "Molecular biology", formation of water-soluble globular protein always starts from amino acid sequence of a protein or protein primary structure, which is specified by gene expression, or transcription and translation of a gene. Polypeptide chain produced by the gene expression is, first, folded into the respective secondary structures (α-helix, β-sheet and turn/coil structures) and, further, the secondary structures are assembled to tertiary structure (Figure 1) [1].

On the other hand, S. Ohno was proposed "gene duplication theory" for new gene formation about 50 years ago. The theory suggests that new gene is produced from one of duplicated genes [2]. Thereafter, it has been confirmed that the theory can well explain the way, how new homologous genes in a family gene have been created. However, the gene duplication theory has a fatal weakness that a previously existing gene is always required to create a new homologous gene. Therefore, the theory cannot show the way, how entirely new gene/protein (EntNew gene/protein) or the first

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 December 2019 doi:10.20944/preprints201912.0021.v1

family gene/protein was created. The word, gene/protein, is used to indicate gene and/or protein in this article, because both are always intimately related to each other after the first gene/protein was created.

It would be obvious that both base sequence and amino acid sequence of mature gene/protein cannot be designed in advance, and, therefore, EntNew gene/protein must be produced by random polymerization of nucleotides and amino acids. However, EntNew gene/protein could not be produced by random joining of the respective monomeric units, because diversities of base sequence and amino acid sequence for a small protein composed of 100 amino acids are extraordinary large as $(4^3)^{100} = \sim 10^{180}$ and $20^{100} = \sim 10^{130}$, respectively (Figure 1) [3]. Furthermore, the genetic information could not be generated without a target of protein. Therefore, it is totally unknown how EntNew gene/protein is created, regardless of significance of the creation of EntNew gene/protein in biological sciences (Figure 1). Those would be the reason, why formation process of protein starts from amino acid sequence in every textbook still now.

I previously described in several papers that a special amino acid composition or protein 0^{th}-order structure is essential to understand the creation of EntNew gene/protein through random polymerization of the respective monomeric units [4-9]. However, the descriptions in the previous papers were rather fragmentary. Then, I describe the characteristics and significance of protein 0^{th}-order structure through the creation mechanism of EntNew gene/protein in this review article. The reason, why the number of references in this article is not many, is because almost no investigation about the creation of EntNew gene/protein has been reported except our studies so far.

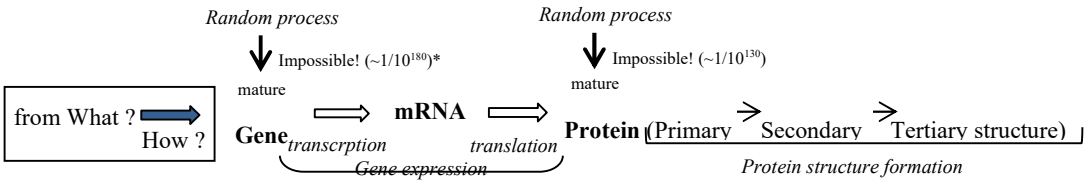


Figure 1. Gene expression (transcription and translation) and protein structure formation. After polypeptide chain was synthesized by the gene expression, the chain (primary structure) is folded into secondary and tertiary structures However, both mature gene and mature protein could not be created directly through random processes. On the other hand, understanding the process of EntNew gene/protein should be one of the most important issues in the field of biological sciences. Nevertheless, it is totally unknown still now, how and from what mature gene/protein was created. Asterisk (*) indicates the result calculated under the genetic code without degeneracy.

2. Creation of EntNew gene/protein under protein 0^{th}-order Structure

2.1. Properties of modern water-soluble globular protein and protein 0^{th}-order structure

It would be unquestionable that six properties (hydrophobicity/hydrophilicity (hydropathy), α-helix, β-sheet, turn/coil formabilities, acidic amino acid (Acidic-AA) content and basic amino acid (Basic-AA) content) are important for protein structure formation. Hence, the six protein structure indexes of water-soluble globular proteins encoded by genomes of seven microorganisms with different GC content (*Mycobacterium tuberculosis* (GC=65.6%), *Aeropyrum pernix* (56.3%), *Escherichia coli* (50.8%), *Bacillus subtilis* (43.5%), *Haemophilus influenzae* (38.1%), *Methanococcus genitarium* (31.3%)

 doi:10.20944/preprints201912.0021.v1

and *Borrelia burgdorferi* (28.2%)) were calculated with amino acid composition in a protein and hydropathy and secondary structure indexes of amino acid given in Stryer's text book [1], according to the following equation for the former four protein indexes [10]. $I(x)_t = \Sigma\ I(x)_a\ n_a/n_t$, where $I(x)_t$, $I(x)_a$, n_a and n_t are total index of a protein, index for each amino acid, each amino acid number and total amino acid number in a protein, respectively. Acidic-AA (aspartic and glutamic acids) and basic-AA (histidine, lysine and arginine) contents were arithmetically obtained. The six average protein structure indexes, which were obtained with water-soluble globular proteins encoded by the seven genomes, were as follows: Hydropathy (-1.51 ± 0.38), α-helix (1.03 ± 0.03), β-sheet (1.00 ± 0.02), turn/coil (0.96 ± 0.05) formabilities, Acidic-AA content (12.0 ± 3.2%) and Basic-AA content (14.1 ± 3.1%) [10]. Therefore, it is expected that if all the six protein structure indexes of an imaginary protein fall within the respective ranges of the six indexes, polypeptide chain of the imaginary protein should be folded into water-soluble globular structure.

There exist three protein 0^{th}-order structures corresponding the respective evolutionary stages of the genetic code. Those are; (1) amino acid composition (actually, amino acid sequence) of imaginary protein encoded by nonstop frame on antisense strand of GC-rich gene (GC-NSF(a)) under the modern genetic code. The amino acid composition is similar to SNS-encoding 10 amino acids [11]. (2) 10 amino acids ([GADV]-amino acids plus Glu, Leu, Pro, His, Gln and Arg) encoded by SNS primitive code. [GADV] means four amino acids, Gly [G], Ala [A], Asp [D] and Val [V]. (3) four [GADV]-amino acids encoded by GNC primeval code.

2.2. The reason why formation of mature gene/protein must start from immature gene/protein

One-dimensional genetic information for even a small protein could not be created by random joining of nucleotides, as can be seen in Figure 1. On the other hand, mature protein has been generally optimized into a rigid structure, so that enzymatic activity of the protein can be restricted into one substrate at one active site. The rigidity of a protein makes it possible to exclude other organic compounds from the catalytic site. Of course, such a refined protein cannot be produced by random joining of amino acids, because one amino acid sequence of the protein cannot be picked up from the extraordinary large sequence diversity, $20^{100} = \sim 10^{130}$ (Figure 1) [3].

There would be only one way, with which such a refined or mature rigid protein can be created. That is to start from immature protein with some flexibility, which is produced by random joining of amino acids in a protein 0^{th}-order structure (Figure 2a) [9]. The reasons are as follows.

(1) Many, probably several hundreds, catalytic sites should appear on surface of even a small globular protein composed of 100 amino acids, taking the combination number of two to four surface amino acids into consideration.

(2) A possible catalytic site on the globular protein with some flexibility could accept even newly encountered organic compound owing to swinging of surface amino acid residues, giving a clue for the immature protein to evolve to mature protein (Figure 2c).

(3) As protein is a polymer composed of amino acids, an imperfect catalytic site of an immature protein can be adjusted to structure of an organic compound through replacements with the variety of amino acids, so that the structure around an active center can be subtly changed every one amino

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 December 2019 doi:10.20944/preprints201912.0021.v1

acid replacement even at a remote place of the protein, as like an allosteric transition of a protein. Thus, mature protein would be formed from immature protein through repetitions of base substitutions accompanied by the subtle adjustments.

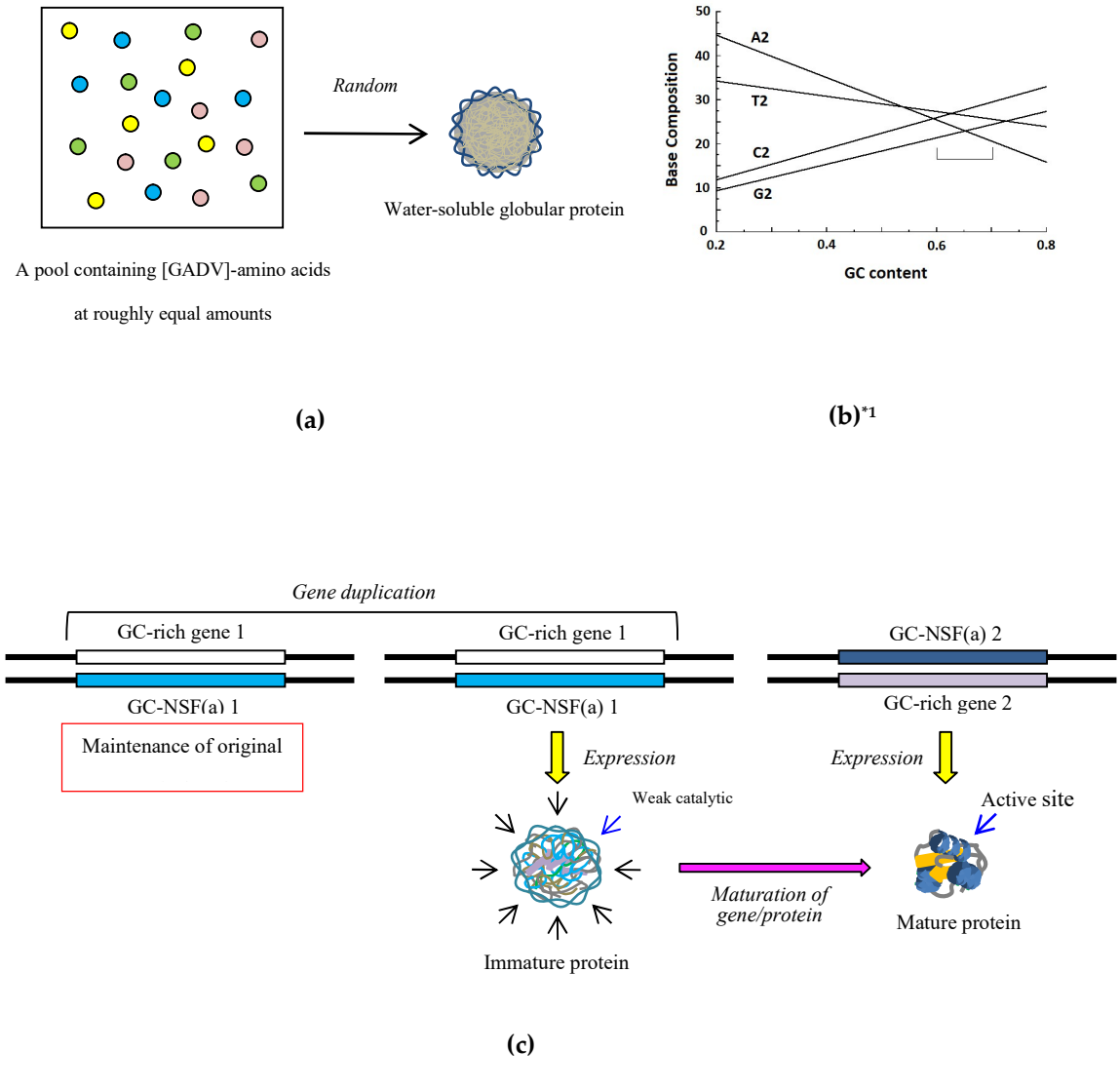


Figure 2. (a) All proteins synthesized by random joining of amino acids, which are stored in a pool, should have the same amino acid composition. Therefore, a polypeptide chain could be folded into water-soluble globular structure, if the polypeptide chain satisfies the six protein structure indexes, which were obtained by calculation with amino acid composition of water-soluble globular proteins, (see also Section 2.1) [7]. Small colored circles represent [GADV]-amino acids and blue wavy lines on the circle symbolize flexible structure of the protein. **(b)** Dependencies of base compositions at second codon position, which were obtained with genomic data of the seven microorganisms (Section 2.1) [4]. Hydrophobicity of imaginary protein encoded by GC-NSF(a) becomes smaller in GC-rich region (more than 60% GC content), because thymine content at the second codon position (T_2) becomes smaller than adenine content (A_2) in the region. Note that T_2 and A_2 on sense strand are replaced with A_2 and T_2 on antisense strands, respectively. **(c)** Mechanism for creation of EntNew gene from codon sequence on GC-NSF(a). EntNew gene/protein could be created from GC-NSF(a) 1,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 December 2019 doi:10.20944/preprints201912.0021.v1

if a low but necessary catalytic activity (thin blue arrow) could be found on a surface of the immature flexible protein produced by gene expression of GC-NSF(a) 1. The reason is because the immature protein could be matured to a protein with a high catalytic activity through accumulation of base substitutions on the GC-NSF(a) necessary to maturation of the protein (bold reddish violet arrow). The creation mechanism of EntNew gene/protein in SNS code era or GNC code era can be similarly understood by replacing GC-rich gene and GC-NSF(a) with $(SNS)_n$ gene or $(GNC)_n$ gene and $(SNS)_n$(a) or $(GNC)_n$(a), respectively.

(4) Next, consider what kind of situation arises, if creation of mature protein starts from a protein with extremely high rigidity. Even a clue for creation of the mature protein cannot be found on any sites on such a too rigid protein, because a possibility, that structure of an organic compound exactly matches the surface structure of the rigid protein, would be zero. Therefore, any organic compound cannot be accepted by the rigid protein to express even a low catalytic activity. This could be easily understood by considering a following example: A key roughly made with metal, never match a keyhole, which was also roughly made with metal.

(5) The substance enabling to produce a mature protein from such an immature protein is to start from a water-soluble globular structure with some flexibility, which is produced by random joining of amino acids in protein 0^{th}-order structure (Figure 2c).

Then, I describe three processes generating an EntNew mature gene/protein under three genetic codes (the universal or standard genetic code, SNS primitive code and GNC primeval code), corresponding to three protein 0^{th}-order structures, so that significance of the protein 0^{th}-order structure could be more deeply understood.

2.3. GC-NSF(a) hypothesis on EntNew gene/protein creation

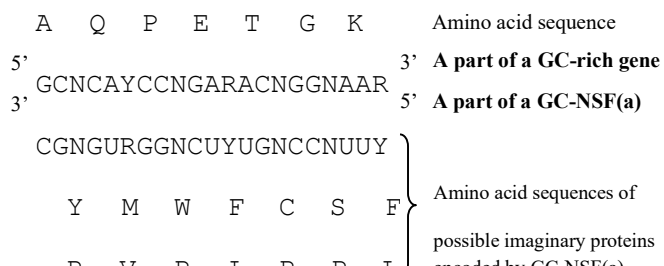
2.3.1. The reason why GC-NSF(a) can encode water-soluble globular protein with some flexibility

The reasons, why EntNew gene/protein is effectively generated from GC-NSF(a), are as follows [11].

(1) Nonstop frame appears on antisense strand of GC-rich gene or GC-NSF(a) at a high probability, because three termination codons, UAA, UAG and UGA, are AU-rich.

(2) Two amino acid sequences of proteins encoded by GC-rich gene and its GC-NSF(a) are quite different from each other (Figure 3), because two strands of DNA are antiparallel and the universal genetic code is triplet and asymmetric. Nevertheless, both amino acid compositions encoded by GC-rich gene and its GC-NSF(a) are similar to $(SNS)_n$ and polypeptide chains produced under the SNS code are folded into water-soluble globular structure at a high probability [4,10].

(3) Structure formabilities of all proteins with the same amino acid composition are the same, because the structure formabilities were obtained by calculation using amino acid composition and structure indexes of the respective amino acids [1,10]. Therefore, all proteins synthesized by random joining of amino acids in a pool should have the same protein structure formabilities. This lead to the idea, protein 0^{th}-order structure (Figure 2a) [4-9].



 doi:10.20944/preprints201912.0021.v1

Figure 3. Amino acid sequences encoded by GC-rich gene and the corresponding codon sequence on GC-NSF(a). The two amino acid sequences are quite different from each other. Possible amino acid sequences of imaginary protein encoded by GC-NSF(a) are also shown, which are generated by base substitutions upon degeneracy at third codon position of the GC-rich gene [12].

(4) Amino acid composition (actually, amino acid sequence) encoded by GC-NSF(a) satisfies the six conditions for water-soluble globular protein formation (see Section 2.1) [6,10]. Therefore, the polypeptide chain produced from GC-NSF(a) could be folded into water-soluble globular structure, which is similar to extant water-soluble globular protein, at a high probability.

(5) Mature protein encoded by extant gene is generally rigid [1]. On the contrary, immature protein produced from GC-NSF(a) should be more flexible than mature protein encoded by GC-rich gene, because of a small hydrophobicity (Figure 2b) and a high Gly content of the protein [11]. The high flexibility makes it possible to adapt surface residues to even never encountered organic compound and to give a clue for the immature protein to evolve to mature protein (Figure 2c).

(6) Taking the number of combination of several tens of surface amino acids into consideration, it is expected that many catalytic sites, probably from several hundreds, should appear on surface of an immature protein. The number of possible catalytic sites should increase much more, owing to flexible structure of the immature protein synthesized by random polymerization of amino acids in protein 0^{th}-order structure, although the catalytic activities would be considerably low. However, the low catalytic activity would be out of question, because the catalytic activity expected must be absent at the time point when the EntNew protein was required. Therefore, mature EntNew protein could be produced by raising the low activity on the surface of the immature protein, if the weak but sufficiently high activity could be found out among many candidates of the catalytic activity on the protein surface (Figure 2c).

(7) Furthermore, degeneracy at the third codon position of GC-NSF(a) could also contribute to generate diverse amino acid sequences without replacing amino acid sequence of a protein encoded by the gene on sense strand. Thus, the degeneracy could also assist the creation of EntNew gene/protein (Figure 3) [12].

Based on the properties of GC-NSF(a) as described above, it can be concluded that GC-NSF(a) effectively encodes one of amino acid sequences, which are produced by random joining of amino acids in the protein 0^{th}-order structure (Figure 2c), and, therefore, mature water-soluble globular protein could be produced from the GC-NSF(a) at a high probability. Direct evidence for the GC-NSF(a) hypothesis has been obtained through homology search between amino acid sequences of modern proteins encoded by GC-rich genes and imaginary proteins encoded by GC-NSF(a)s on the genome of *Pseudomonas aeruginosa* PAO1 [13].

2.3.2. Formation process of EntNew gene/protein under the universal (standard) genetic code

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 December 2019 doi:10.20944/preprints201912.0021.v1

Here, discuss the formation process from immature gene/protein to mature gene/protein under the universal (standard) genetic code (Figure 2c) [9].

(1) First, a GC-rich gene 1 is duplicated into two genes.

(2) Original genetic function is conserved by one of the duplicated genes.

(3) An immature protein is produced by expression of the GC-NSF(a) 1 on antisense strand of the other GC-rich gene 1 (Figure 2c).

(4) Many possible catalytic sites would appear on surface of the immature flexible protein, because the number of combination should reach to several hundreds. Swinging of the surface residues on the protein further increases the effective number.

(5) If one catalytic function can be detected from many candidates for a catalytic activity on the immature protein 1, the catalytic activity could be elevated to a level of sufficiently high activity by receiving multiple base substitutions necessary to raise the activity (Figure 2c).

(6) In parallel, structure of the immature protein gradually increases rigidity to refine the catalytic activity and to be able to reject any other organic compounds.

(7) Simultaneously, original mature gene 1 on sense strand and immature GC-NSF(a) gene 1 on antisense strand become immature GC-NSF(a) gene 2 and mature gene 2, respectively (Figure 2c).

It is important to understand that all catalytic activities, which are necessary for extant organisms to live, could be found somewhere on surfaces of a number of immature globular proteins, which are produced from many GC-NSF(a)s, although the activities at initial evolutionary stage might be, of course, extremely low. All extant organisms are living on this planet using EntNew genes/proteins and their homologous progeny genes/proteins, which were derived from the first family gene/protein. Inversely stating that, any creature could not evolve to modern organisms on the present Earth, if the mechanism were not acquired, because both the first EntNew family gene/protein and homologous genes/proteins in the family could not be acquired.

2.4. Creation of EntNew gene/protein under SNS primitive genetic code

Mechanism generating EntNew gene/protein under SNS genetic code encoding 10 amino acids is quite similar to that under the universal genetic code (Figure 2c) [10,14]. Therefore, the mechanism used in the SNS code era is described in this section, focusing on only main points, because their details are described in the previous Section 2.3.

It is naturally considered that gene used in the SNS code era was SNS repeating sequence or $(SNS)_n$. Under the SNS genetic code, EntNew gene/protein could be generated from nonstop frame on antisense strand of $(SNS)_n$ gene ($(SNS)_n(a)$), because $(SNS)_n(a)$ could code for a random amino acid sequence and immature water-soluble globular protein with some flexibility could be produced from the $(SNS)_n(a)$, as similarly as the case of modern GC-NSF(a) (Figure 2c). Many possible catalytic sites should appear on the surface of the immature protein, which was synthesized by expression of $(SNS)_n(a)$. The reasons, why $(SNS)_n(a)$ effectively can encode one of random amino acid sequences synthesized in a protein 0^{th}-order structure or 10 amino acids encoded by SNS code, are as follows.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 December 2019 doi:10.20944/preprints201912.0021.v1

Amino acid composition (actually amino acid sequence) encoded by a $(SNS)_n$ gene satisfies the six conditions for water-soluble globular protein formation (Figure 4a). Two amino acid sequences encoded by $(SNS)_n$ gene and by $(SNS)_n(a)$, are quite different from each other, similarly to the case of GC-rich gene and GC-NSF(a). Therefore, it can be concluded that $(SNS)_n(a)$ effectively encodes one of amino acid sequences, which were produced by random joining of 10 amino acids. In addition, immature protein, which is produced from $(SNS)_n(a)$, should have a small hydrophobicity causing some flexibility necessary to adjust amino acid residues on the protein to a new organic compound (Figure 2c), although the catalytic activity would be quite low. In addition, degeneracy, G or C, at the third codon position enlarged the range of choice of amino acid sequence to create EntNew gene/protein. Those are also similar to the cases of GC-NSF(a) of extant GC-rich gene.

Therefore, formation process of EntNew gene/protein can be explained by replacing GC-rich gene and GC-NSF(a) with $(SNS)_n$ gene and $(SNS)_n(a)$, respectively (Figure 2c).

(1) First, one $(SNS)_n$ gene 1 was duplicated into two. Original genetic function was conserved in one of the duplicated genes. Therefore, multiple mutations could be accepted on the other duplicated $(SNS)_n(a)$ 1 to create EntNew gene (see also Figure 2c).

(2) An immature protein 1 could be produced by expression of the $(SNS)_n(a)$ 1 (Figure 2c).

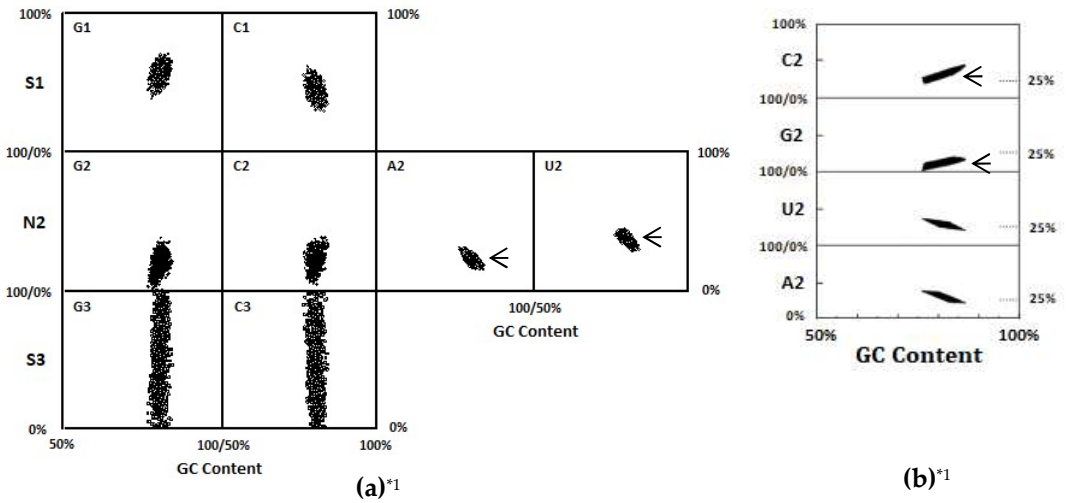


Figure 4. (a) Dot plot of base compositions at three codon positions of mature proteins encoded by SNS code which satisfies the six conditions for protein structure formation, which were obtained as average properties of extant proteins encoded by genomes of seven micro-organisms [10]. **(b)** Dot plot of base compositions at the second codon position of mature proteins which satisfies the four conditions, which were obtained by excluding two conditions, Acidic-AA content and Basic-AA content from the six conditions for protein structure formation [10]. Small arrows indicate that base contents at the respective base positions are less or more than the corresponding base contents.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 December 2019 doi:10.20944/preprints201912.0021.v1

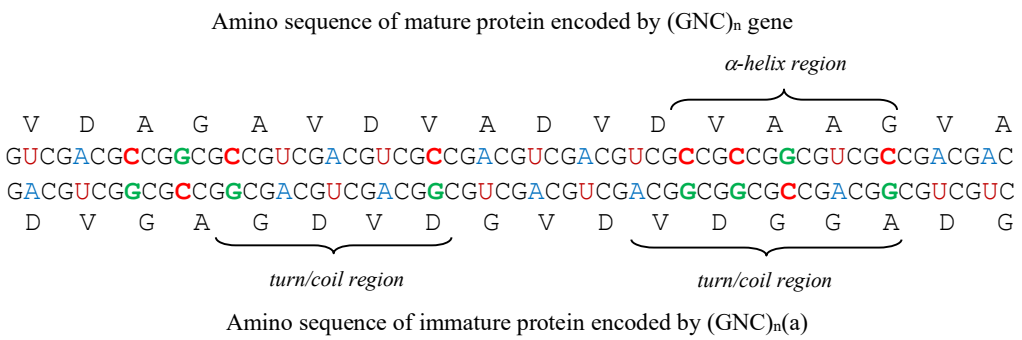
(3) If one catalytic function was found on the surface of the immature protein, the catalytic activity could be elevated as receiving multiple base substitutions (see also Figure 2c). In parallel, structure of the immature protein gradually became rigid and compact to grow to mature protein. Simultaneously, the original mature gene 1 and the immature $(SNS)_n$(a) 1 became immature $(SNS)_n$(a) 2 and mature $(SNS)_n$ gene 2, respectively (Figure 2c).

2.5. Creation of EntNew gene/protein under GNC primitive genetic code

It is explained in more detail in this section, how EntNew $(GNC)_n$ gene encoding [GADV]-protein was generated from GNC codon sequence on its antisense strand, $(GNC)_n$(a), because it should be easier to understand the evolutionary process from $(GNC)_n$(a) than the above cases, as GNC primeval genetic code encodes only four [GADV]-amino acids [10,15].

One of protein 0^{th}-order structures is the amino acid composition, in which [GADV]-amino acids are contained at roughly equal amounts, because protein with the amino acid composition well satisfies the four conditions, which are obtained by exclusion of both Acidic-AA content and Basic-AA content from the six conditions (Figure 4b). The reason, why Acidic-AA content and Basic-AA content can be excluded from the six conditions, is because Acidic-AA and Basic-AA could be replaced by divalent anions, as CO_3^{2-}, SO_4^{2-}, and divalent cations, as Mg^{2+}, Mn^{2+}, respectively. As a mature protein like a precision molecular machine never be directly produced by random joining of [GADV]-amino acids in the absence of genetic function, such a mature protein must be generated from an immature flexible water-soluble globular protein encoded by premature gene carrying random GNC codon sequence, followed by evolutionary process [9].

[GADV]-protein produced from $(GNC)_n$ gene satisfies the four conditions, when Ala encoded by GCC codon was used more frequently than Gly encoded by GGC codon in the protein (Figure 4b). In other words, the protein could take the structure similar to extant proteins, when α-helix forming amino acid, Ala, was used considerably more than turn/coil forming amino acid, Gly. This indicates that immature [GADV]-protein synthesized from $(GNC)_n$(a) must be more flexible than mature [GADV]-protein, because GCC codons for Ala on sense strand are replaced by GGC codons for Gly on antisense strand. Therefore, contents of Ala and Gly in mature protein encoded by a gene on sense strand, are equal to Gly and Ala contents in the corresponding immature protein encoded by $(GNC)_n$(a), respectively (Figure 5). It means that Gly was used at a higher frequency than Ala in the immature protein produced from $(GNC)_n$(a) and, therefore, the structure of the imaginary protein became flexible (Figure 5). The flexible structure enabled a premature catalytic site of the protein to easily adjust to a newly encountered organic compound.



 doi:10.20944/preprints201912.0021.v1

Figure 5. Amino acid sequences of mature protein encoded by $(GNC)_n$ gene (upper base sequence) and of immature protein encoded by $(GNC)_n(a)$ (lower base sequence). Letters written with colored alphabets indicate four bases, green bold G: guanine, red bold C: cytosine, blue A: adenine, and brown U: uracil. Six Cs (30%), As (30%), Us (30%) and two Gs (10%) are randomly arranged at the second codon position of an $(GNC)_n$ gene, taking the results of computer analysis (C_2 (31%), G_2 (11%), U_2 (29%) and A_2 (29%)) shown in Figure 4b into consideration [10]. It can be seen that two amino acid sequences encoded by $(GNC)_n$ gene and $(GNC)_n(a)$ are quite different from each other. The difference is caused by anti-parallel double-stranded RNA structure and asymmetric GNC code. Consequently, polypeptide chain encoded by codon sequence on antisense strand is folded into turn/coil conformation more frequently than that encoded by sense codon sequence.

Formation process of EntNew gene/protein in GNC code era is explained as follows (see also Figure 2c), as the process is basically the same under the three genetic codes, GNC code, SNS code and the universal genetic code.

(1) Multiple mutations could be introduced on $(GNC)_n(a)$ sequence to create EntNew $(GNC)_n$ gene, because original genetic function could be conserved in surplus one of the duplicated $(GNC)_n$ genes, after gene duplication (Figure 2c). If a catalytic activity, which was required to live, could be found at one site on the surface of the immature protein 1, the catalytic activity could evolve to a sufficiently high level upon amino acid replacements necessary to raise the activity (Figure 2c).

(2) Upon the evolution of the gene/protein, the immature protein became rigid to be able to refine the protein structure and reject any other organic compound. Simultaneously, the original mature $(GNC)_n$ gene 1 and the immature $(GNC)_n(a)$ gene 1 became immature $(GNC)_n$ gene 2 and mature $(GNC)_n$ gene 2, respectively (see also Figure 2c).

Thus, the excellent assignment of four [GADV]-amino acids into four GNC codons made it possible to create EntNew $(GNC)_n$ gene/[GADV]-protein efficiently.

3. Discussion

One of the important problems in the field of biological sciences, how and from where EntNew gene/protein has been created, remains unsolved still now (Figure 1). The reason is because there are several matters, which make it difficult to solve, as described below.

(1) Genetic information for protein synthesis is written as one-dimensional base sequence. On the contrary, catalytic function is expressed on a surface of a water-soluble globular protein with three-dimensional structure.

(2) As a matter of course, it is impossible to design in advance the first family gene/protein, which is totally different from any previously existing genes/proteins, and, therefore, does not show any meaningful sequence homology with them.

(3) Therefore, such an EntNew gene/protein must be created through random process. However, it would be impossible to create such a gene/protein with ordered sequence through

 doi:10.20944/preprints201912.0021.v1

367 random processes, because of a quite high wall of extraordinary sequence diversities for creation of
368 EntNew gene/protein.

369 (4) On the other hand, it would be also impossible to create EntNew gene independently of
370 protein, because gene codes for amino acid sequence of protein. Of course, any mature protein with
371 a specific amino acid sequence also cannot be created independently of gene, because the gene is
372 directly connected with amino acid sequence of the protein.

373 Taking the contradictory request for creation of mature EntNew gene/protein into
374 consideration, only one possible way would be to start from immature protein produced by random
375 polymerization of amino acids in a specific amino acid composition or protein 0^{th}-order structure,
376 which is encoded by one of three GC-rich codon sequences, GC-NSF(a), $(SNS)_n$(a) or $(GNC)_n$(a),
377 because the immature water-soluble globular protein could be evolved to mature protein through
378 base replacements of the antisense sequence, using memory capacity of double-stranded RNA or
379 DNA.

380 $(GNC)_n$(a) could overcome the quite difficulties by introduction of protein 0^{th}-order structure
381 holding the key solving the problem for the first time (Section 2.5). $(SNS)_n$(a) under SNS primitive
382 genetic code and GC-NSF(a) under the universal genetic code succeeded the excellent properties
383 acquired by $(GNC)_n$(a) (Figure 6).

384 That is supported by the facts described below.

385 (1) Base compositions at three codon positions of GC-NSF(a) is comparatively similar to SNS,
386 which encodes water-soluble globular protein at a high probability (Figure 5a) [4,10].

387 (2) Frequency using G-start codons is sufficiently higher than those of C-, A- and U-start codons,
388 in the cases of not only GC-rich genes but also AT-rich genes, indicating that [GADV+E]-amino acids
389 form the basis of protein and, therefore, are used in protein at a high frequency, irrespective of GC
390 content of the gene [4,11]. This also supports that modern genes are successors of the most ancestral
391 $(GNC)_n$ gene (Figure 6).

392

393

394

Completion of

The Universal Genetic Code

C	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	G
G	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	G

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 December 2019 doi:10.20944/preprints201912.0021.v1

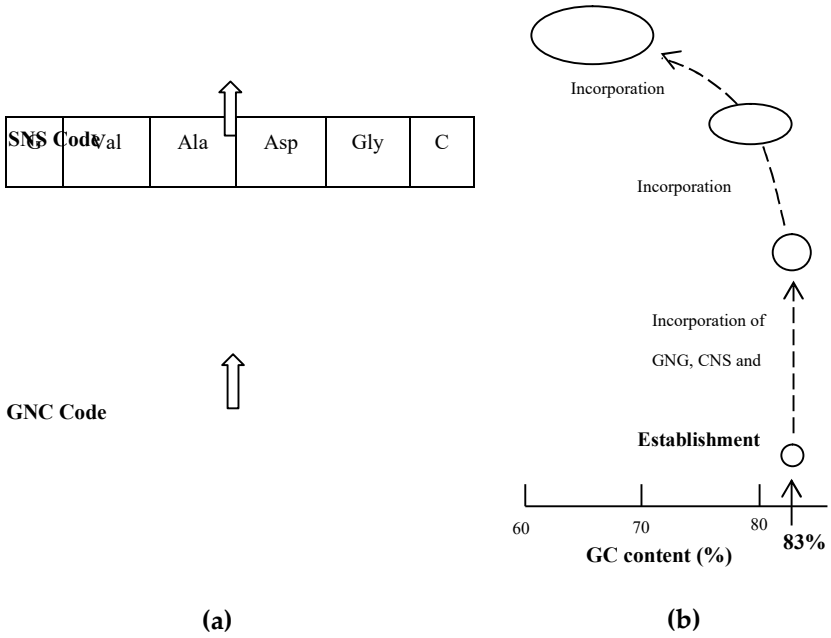


Figure 6. (a) Evolution of genetic code from GNC primeval code to the universal genetic code through SNS primitive code. The evolutionary process is drawn according to GNC-SNS primitive genetic code hypothesis, assuming that the genetic code originated from GNC code encoding four [GADV]-amino acids [10]. **(b)** Change of the field for EntNew gene creation, accompanied by the genetic code evolution. After establishment of GNC primeval genetic code, codons were incorporated into the GNC code in order of GNG, CNS, GNW, CNW, A-start and U-start codons [16]. Upon the evolution of the genetic code, the field for EntNew gene creation shifted from extremely GC-rich (around 83%) to ordinary GC-rich region around 60 to 70%. Circles and ellipsoids indicate the supposed fields of EntNew gene creation.

(3) As shown in Figure 6b, average GC content of $(GNC)_n$ genes used under the GNC primeval genetic code, was quite high as about 83%. GC content of $(SNS)_n$ genes derived from the GC-rich $(GNC)_n$ genes was also high as about 83%. After establishment of SNN primitive code through degeneracy at the third codon position, new codons were further incorporated into SNN code in order of A-start and U-start codons (Figure 6a) [16]. Consequently, GC content of genes used under the universal genetic code has expanded from about 25% to 75%. However, codon sequence on antisense strand expanding from about 60% to 70% GC content remains remnants of extremely GC-rich $(SNS)_n$ (~83%) and $(GNC)_n$ (~83%) genes and has been used as the field for EntNew gene creation still now (Figure 6b) [11].

(4) In GNC primeval genetic code era, maturation of EntNew gene from immature gene was adjusted by exchanging Gly for Ala encoded by GGC and GCC, respectively (Figure 4b). On the contrary, in SNS code era, hydrophilic amino acids encoded by SAS codons were predominantly exchanged for hydrophobic amino acids encoded by SUS codons on evolutionary process from immature to mature protein (Figure 4a).

(5) On the other hand, shift from immature to mature protein under the universal genetic code has been controlled by the two adjustments, one is the exchange of Gly for Ala and the other is replacement of hydrophilic amino acids with hydrophobic amino acids. These also indicate that

 doi:10.20944/preprints201912.0021.v1

modern GC-rich genes and GC-NSF(a)s have been propagated from the most ancestral $(GNC)_n$ gene through $(SNS)_n$ gene.

On the contrary, EntNew gene/protein cannot be produced from antisense sequence of AT-rich gene. The reason is easily explained as follows.

(1) Stop codons appear on the antisense strand of AT-rich gene at a high frequency, because all the three termination codons, UAA, UAG and UGA, are AU- or AT-rich. This means that only short peptides should be synthesized from antisense codon sequence of AT-rich gene.

(2) Imaginary protein encoded by codon sequence on antisense strand of AT-rich gene should be highly hydrophobic, making the protein insoluble in water.

(3) The imaginary proteins have excess β-sheet and extremely low turn/coil formabilities and too low Acidic-AA composition [4]. This indicates that imaginary protein, which is produced from codon sequence on antisense strand of AT-rich gene, does not satisfy the six conditions, indicating that the protein encoded by the AT-rich codon sequence cannot form water-soluble globular structure as a prerequisite for proteinous catalyst.

As described so far, EntNew genes/proteins were and have been created from GC-rich codon sequence on antisense strand, irrespective of any genetic code era, GNC code, SNS code and the universal genetic code. The following properties of the genetic code enabled the efficient creation of EntNew gene/protein.

(1) Hydrophobic amino acids and hydrophilic amino acids are arranged in U_2 and A_2 columns of the universal genetic code, respectively.

(2) In U_2 column, β-sheet (Val, Ile, Phe) and α-helix (Leu, Met) forming hydrophobic amino acids are arranged. On the other hand, turn/coil (Asn, Asp) and α-helix (Glu, Gln, Lys) forming hydrophilic amino acids are arranged in A_2 column. The arrangements of amino acids in the universal genetic code table enable secondary structure formabilities not to be biased in modern proteins, irrespective of GC content of gene.

(3) Protein 0^{th}-order structures, amino acids encoded by GNC code and SNS code, have written in one and four columns of the universal genetic code table, respectively, although it is natural because creation of EntNew gene/protein and protein synthesis were and have been carried out based on the genetic codes.

Such a stunning arrangement of amino acids in the universal genetic code table would be the results, that appropriate amino acids were introduced and captured during evolution of the genetic code. The universal genetic code, which has been formed through evolution from the first GNC genetic code through SNS code [10], made it possible to efficiently create EntNew genes/proteins and for diverse organisms to prosper on the present Earth. I marvel at the beautiful life system, which has been acquired during the evolution.

Funding: This research received no external funding.

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 the origin and evolution of the fundamental life system.

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

References

1. Berg, J.M.; Tymoczko, J.L.; Stryer, L. *Biochemistry*, 5th ed.; W.H. Freeman Comp., New York, U.S.A. 2002.
2. Ohno, S. *Evolution by gene duplication*, Springer, Heiderberg, Germany, 1970.
3. Dill, K.A. Dominant forces in protein folding. *Biochemistry* **1990**, *29*, 7133–7155.
4. Ikehara, K. Origins of gene, genetic code, protein and life: comprehensive view of life systems from a GNC-SNS primitive genetic code hypothesis (a modified English version of the paper appeared in *Viva Origino*, Vol. 29, 66-85 (2001)). *J. Biosci.*, **2002**, *27*, 165-186.
5. Ikehara, K. Possible Steps to the Emergence of Life: The [GADV]-Protein World Hypothesis. *The Chemical Record*, **2005**, *5*, 107-118.
6. Ikehara, K. [GADV]-Protein World Hypothesis on the Origin of Life. "Genesis: In the beginning: precursors of life, Chemical Models and Early Biological Evolution". Seckbach, J. (ed.) Springer, pp. 107-122, 2012.
7. Ikehara, K. Pseudo-replication of [GADV]-proteins and origin of life. *Int. J. Mol. Sci.* **2009**, *10*, 1525-1537.
8. Ikehara, K. [GADV]-protein world hypothesis on the origin of life. *Orig. Life Evol. Biosph.*, **2014**, *44*, 299–302.
9. 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.
10. 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.
11. 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.
12. Ikehara, K. Degeneracy of the genetic code has played an important role in evolution of organisms. *SOJ Genet. Sci.*, **2016**, *3*, 1-3.
13. Oi, R.; Ikehara, K. Direct evidence for GC-NSF(a) hypothesis on creation of entirely new gene/protein. *Curr. Proteom.*, **2018**, *3*, 13-23.
14. Ikehara, K.; Yoshida, S. SNS hypothesis on the origin of the genetic code. *Viva Origino*, **1998**, *26*, 301-310.
15. Ikehara, K. Mechanism for creation of "original ancestor genes". *J. Biol. Macromol.*, **2005**, *5*, 21-30.
16. Ikehara, K. The Origin of tRNA deduced from *Pseudomonas aeruginosa* 5'anticodon-stem sequence -anticodon-stem loop hypothesis-. *Orig. Life Evol. Biosph.* **2019**, *49*, 61-75. doi. org/10.1007/s11084-019-09573-w