 doi:10.20944/preprints202103.0003.v1

Article

Proposition of a Novel Strategy for Creation of Entirely New Proteins

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: Proteins having a variety of functions play many essential roles in maintaining various life activities in organisms. Various methods, by which new protein functions can be artificially produced, have progressed rapidly upon development in recombinant DNA technology and effective screening techniques. However, the obtainable scope of the new functions has been restricted in a narrow range, because only functions of presently existing proteins can be used. On the other hand, it has been considered that it would be impossible to create an entirely new protein, which does not show any meaningful homology with any other amino acid sequences of previously existing proteins. The reason is because one amino acid sequence for a protein cannot be selected out from an extraordinary large amino acid sequence diversity as ~10^{130}. As a matter of course, it is impossible to design an amino acid sequence of a protein in advance and a gene encoding the protein cannot be also formed through random process. Nevertheless, extant organisms have generated a variety of entirely new proteins in some way to make full use of them. This means that extant organisms have equipped a mechanism with which entirely new proteins can be produced under the present core life system composed of protein, tRNA (genetic code) and gene. In this article, first I introduce the mechanism, with which entirely new proteins are created in extant organisms, and further propose a novel strategy for application of the mechanism to protein engineering through creation of entirely new proteins, which could contribute to development of various industries.

Keywords: protein engineering; creation of an entirely new protein; pluripotency of an immature protein; GC-NSF(a) hypothesis; protein 0^{th}-order structure; origin of protein

1. Introduction

Diverse and splendid organisms are inhabiting on the present Earth. Such prosperity of terestrial lives have been achieved by evolution during about 4 billion years. Even microorganisms as *Escherichia coli* are living with several thousands of mature proteins. Mammals including human beings, are using several tens thousands of mature proteins to live. In addition, every contemporary protein or a mature protein having a splendid structure and function is produced through genetic expression under the present genetic system and supports many life activities in organisms (Figure 1). Here, it should be noted that proteins mainly support the life activities and, on the other hand, genes play only auxiliary roles in producing proteins with amazing functions [1]. Note also that protein means generally a mature one, which is optimized to like as a precision polymer machine under the genetic system. A term "a mature protein" is used in this article to discriminate from "an immature protein", which is produced through a random process

 doi:10.20944/preprints202103.0003.v1

in the absence of gene or a kind of random process accompanied by expression of a nonstop frame on antisense strand of GC-rich gene (GC-NSF(a)) [2,3]. Therefore, an immature protein is loosely folded into a water-soluble globular structure with some flexibility before optimization (Table 1).

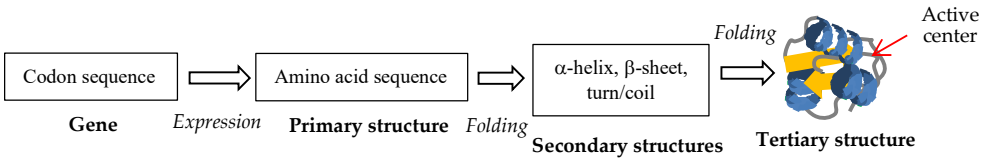

Figure 1. Synthetic pathway of a mature protein. First, a polypetide chain (primary structure) is synthesized according to genetic information for the protein. The polypeptide chain is folded into the respective secondary structures or α-helix, β-sheet and turn/coil structures. The secondary structures are further assembled around a hydrophobic core to form a rigid water-soluble globular protein or tertiary structure carrying one catalytic site usually (Table 1). Thus, mature proteins optimized sufficiently are formed owing to the genetic information.

It has been considered that various protein functions found in extant organisms could be applied to many industrial activities. In facts, many enzyme variants with an improved catalytic activity have been generated with protein engineering or such as, site-directed mutagenesis, metabolic engineering and chemical modification thus far to retain the activity at an elevated temperature and to exhibit a sufficiently high function even in the presence of organic solvents, salts and pH values far from physiological conditions [4,5]. In addition, a new method combining a plural domains paving the ways to generate new protein assemblies [6,7] and evolutionary engineering [8,9] were also invented to construct enzyme variants when lack of structural information impedes the use of rational design. Therefore, it has been expected that the functions could be applied to development of various industries, if new desirable protein functions could be artificially produced with protein engineering. Thus, a variety of successful bioprocesses have been established with the various approaches of protein engineering. However, proteins, which can be used for protein engineering, have been naturally restricted only in presently existing proteins. Therefore, this has become a large obstacle in development of protein engineering.

On the contrary, the restriction could be overcome, if a mechanism, how entirely new (EntNew) proteins with a new function are created in modern organisms, can be understood and the functions of artificially created EntNew proteins can be used as a novel strategy for development of protein engineering. If it becomes possible, the range of protein engineering should greatly expand through the usage of the EntNew proteins constructed artificially. Note that the term, “a new protein”, means a homologous protein, which has been newly formed through duplication of an ancestor gene. On the other hand, “an EntNew protein” means literally “an EntNew one”, which does not show any meaningful homology with any other proteins.

What becomes possible by application of the knowledge for the creation of EntNew proteins to protein engineering?

1. It would become possible to apply the EntNew protein with a new function to synthesis of desirable organic compounds or degradation of unnecessary organic compounds in various industries.
2. It is supposed that function of a new protein, which is constructed by a combination of previously existed protein domains, should be generally low. At such time, it is expected that a mature protein with a high catalytic activity can be acquired using the knowledge and the techniques for the creation of EntNew proteins or through optimization of

 doi:10.20944/preprints202103.0003.v1

the incomplete protein with a weak catalytic function, which was constructed with previously existing techniques for protein engineering.

Then, in the next Section, I first explain the way how modern proteins are produced under the genetic system of extant organisms as a review. After that, I will discuss how EntNew proteins have been created in modern organisms.

Table 1. Comparison of Properties of Mature Protein and Immature Protein

* means the power operator.

	Mature Protein	Immature Protein
1. No. of Gene	1	1 (GC-NSF(a))
2. No. of aa-Sequence	1	1
3. No. of Catalytic Center	1	> 10*130
4. Protein Structure	compact; rigid	swelled; flexible
5. Creation through Random Process	impossible	easy

2. Protein synthesis under the modern genetic system

As well known, proteins are always produced under the modern genetic system composed of three or four main members or gene, tRNA (genetic code) and protein. Therefore, it has been generally considered that a protein as like as a precision machine would never be synthesized in the absence of the corresponding gene (Figure 1). That is, protein synthesis has been considered under the "gene-centered idea" thus far [10]. Many researchers also have considered under the prejudiced idea that gene must be naturally present not only when proteins are synthesized in extant organisms but also even when EntNew proteins are created.

Then, consider in the next Section how genes encoding an EntNew protein have been created in modern organisms.

2.1. An EntNew gene encoding a mature protein never be formed through random process

It would be impossible to create any gene encoding an objective mature protein through random process, because gene or genetic information cannot be formed in the absence of the object, protein, because genetic information for a protein is always formed as referring to a catalytic function of the protein [1]. As a matter of course, nucleotide sequence of a gene encoding a mature protein cannot be designed in advance. Therefore, the gene must be created through a random process. However, a base or a codon sequence must be selected out from an extraordinary vast nucleotide sequence diversity, as $(4^3)^{100}$ = ~10^{180}, even in the case of a gene encoding a small protein composed of 100 amino acids [1]. This means that it is actually impossible to select out one nucleotide sequence for the protein synthesis from the sequence diversity, and that any gene carrying genetic information cannot be actually formed by random joining of nucleotides or codons (Figure 2).

2.2. An EntNew mature protein also never be formed through random joining of amino acids

Then, next consider whether or not an EntNew protein can be created in the absence of gene or through a random process. Any EntNew mature protein also could not be formed through the direct random process (Table 1). The reason is as follows.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2021 doi:10.20944/preprints202103.0003.v1

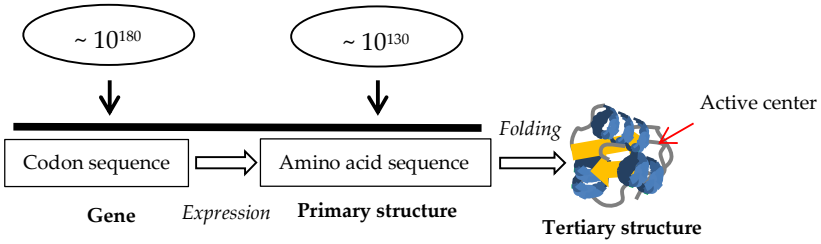

Figure 2. A mature protein is synthesized under genetic system through polypeptide synthesis as described in Figure 1. Therefore, an EntNew protein could be produced, if a codon sequence encoding the amino acid sequence and an amino acid sequence itself can be formed by random joining of the respective monomeric units, codons or amino acids. However, it is impossible to synthesize both the codon sequence and the amino acid sequence through a random process, because of the respective extraordinary vast sequence diversity as ~10^{180} or ~10^{130}.

As similarly to the case of the EntNew gene creation, an amino acid sequence of a specified mature protein cannot be designed previously. Accordingly, any EntNew protein must be created by random joining of amino acids. However, it is also actually impossible to form an EntNew mature protein through direct random joining of amino acids, because one amino acid sequence for formation of a mature protein cannot be selected out from the extraordinary large amino acid sequence diversity or about 20^{100} = ~10^{130} (Figure 2) [11].

Nevertheless, various organisms are inhabiting on the present Earth, as generating in some way and making full use of many EntNew proteins. This means that extant organisms have a mechanism, with which EntNew proteins can be produced under the core life system composed of protein, tRNA (genetic code) and gene (Figure 1), and that terrstrial lives acquired a skillful strategy for creating such mature EntNew proteins about 4 billion years ago. However, it has been totally unknown how EntNew proteins have been created even in extant organisms. The fact, that no paper describing the creation mechanism of EntNew protein has been published until now except my idea, GC-NSF(a) hypothesis [2,3], clearly indicates that.

3. The creation mechanism of EntNew protein in modern organism

I explain the mechanism, how EntNew proteins have been created in extant organisms, in order.

3.1. Various EntNew proteins have been created in extant organisms anytime when necessary

It is well known that there exist many proteins, which have not meaningful homology with any other groups of proteins. Those are classified as the respective protein families [12]. Existence of such protein families clearly indicates that the protein family was not derived from any proteins belonging to the other protein families, meaning that an ancestor protein of the family was generated under a mechanism creating EntNew proteins or independently of any other proteins, and that there should exist the mechanism creating EntNew proteins through an exactly or a kind of random process. After that, many homologous proteins could be produced from the EntNew protein as an ancestor through gene duplication to form one protein family.

3.2. Contradiction between the fact that EntNew proteins have been created in extant organisms and the requirement that the EntNew proteins must be created through a random process

Addressing the point at hand, it is impossible to use any amino acid sequence of previously existing proteins to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2021 doi:10.20944/preprints202103.0003.v1

create a EntNew mature protein, as described in Section 2.2. On the other hand, EntNew mature proteins should have been created through a random process in some way. However, any amino acid sequence of mature proteins cannot be designed beforehand. In addition, it is impossible to create a mature protein like a delicate polymer machine at one stroke through random process on the primitive Earth or even in extant organisms, as repeatedly described. Namely, there is contradiction between the fact that EntNew proteins have been created in extant organisms and the requirement that the EntNew proteins must be created through a random process (Figure 3) [1].

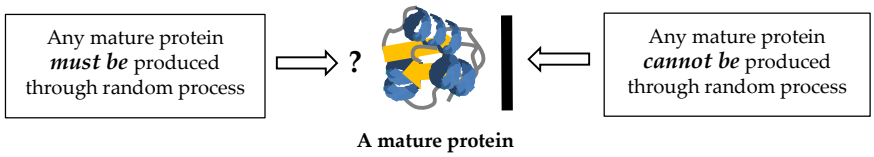

Figure 3. There is contradiction, which it is quite difficult to overcome, in creation process of an EntNew protein. One is that an EntNew mature protein ***must be*** produced through random process and the other is that it ***is impossible*** to produce a mature protein through random process.

3.3. The one way with which an EntNew protein can be created

The one way, how the contradiction between the two requests: (1) an EntNew mature protein must be created through at least a kind of random process, (2) on the other hand, it is impossible to create an EntNew mature protein through random process, can be avoided, is to create indirectly a mature protein using a something unknown, which is produced through a random process (Figure 4). Actually, the one way for overcoming the discrepancy is to first synthesize an immature protein with not exactly but essentially random amino acid sequence and, successively, to optimize gradually the immature protein as referring the catalytic activity to a mature protein as using the ability of double-stranded DNA to memorize for amino acid replacements (Figure 5).

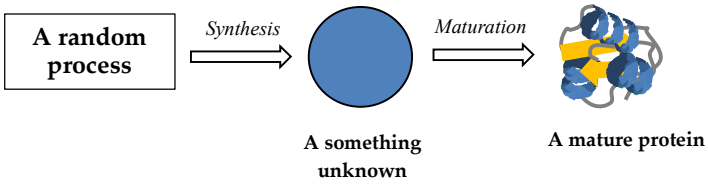

Figure 4. Probably, the only one way for overcoming the discrepancy between (1) the mature protein must be created through random process and (2) a mature protein cannot be produced through a random process, is to produce a something unknown through a random process, and thereafter the something is transformed to a mature protein through maturation process.

3.4. Five factors making it possible to create an EntNew protein

Then, explain the way how a mature protein can be generated through a kind of random process. For the purpose, it must be understood that the following five factors are the keys, which make it possible to create EntNew proteins.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2021 doi:10.20944/preprints202103.0003.v1

1. The first one is protein 0th-order structure, under which immature but water-soluble globular proteins can be produced through a random process (Table 2). It is possible to produce immature proteins by a direct random joining of amino acids in a protein 0th-order structure or [GADV]-amino acids [3,13-15], where [GADV] means four amino acids; Gly [G]; Ala [A]; Asp [D]; and Val [V]. However, it would be impossible to evolve the immature protein produced by such a random process to a mature protein, because amino acid replacements favorable for maturation cannot be memorize in the absence of double-stranded DNA.

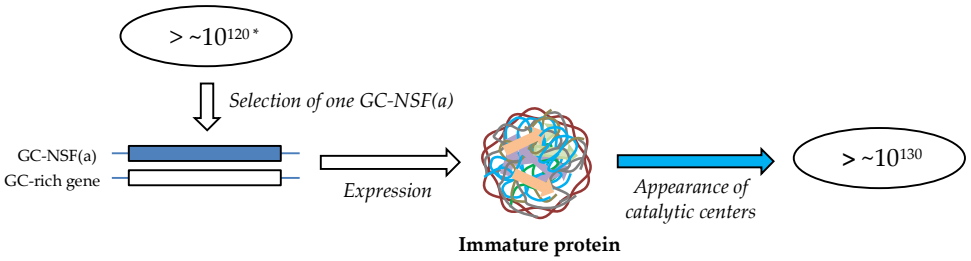

Figure 5. Pluripotency of an immature protein with some flexibility is synthesized through expression of a nonstop frame on antisense strand of GC-rich gene (GC-NSF(a)) [2,3]. It is assumed that catalytic centers more than 10^{130} could be appeared on the surface of one immature protein owing to the flexibility of the immature protein (see the text). Asterisk (*) in the left upper ellipse means that the number 10^{120} is estimated as the minimum base sequence diversity, $(4^2)^{100}$ = ~10^{120}, which is obtained when the complete degeneracy at the third codon position is assumed for simplicity.

2. Therefore, the second one is a double-stranded DNA memorizing amino acid replacements necessary to evolve the immature protein to a mature one. For the purpose, it is necessary to use an essentially random amino acid sequence encoded by a double-stranded DNA or GC-NSF(a). Hereby, it would be possible to use an ability of the double-stranded DNA for memorizing amino acid replacements and to evolve the immature protein to a mature protein (Figure 5) [1].

3. The third one is pluripotency of the immature protein generating many possible catalytic sites on the surface of the protein with some flexibility [1]. Many various surface structures, which are formed especially by wobbling of surface amino acids, appear on one immature protein (Figure 5). One of the respective structures formed by the wobbling corresponds to one mature protein, on which only one catalytic site is formed. Thus, a weak but necessary catalytic activity to live could be detected at a high probability. Actually, an immature protein with some flexibility exhibits diverse catalytic sites or pluripotency on the surface of the protein (Table 2). The pluripotency makes it possible to take the first step to evolution from an immature protein to a mature protein. Thus, EntNew proteins have been created any time when necessary, as if something had previously known the mechanism for creation of such mature proteins.

It might be difficult for many persons to consider that protein structures as more than 10^{130} could be formed through the wobbling of amino acid residues of one immature protein (Figure 5). However, the fact, that many EntNew proteins encoded by the respective genes have been constantly created during evolution of organisms, supports the possibility. The number of 10^{130} is estimated from calculation of $100^{\sim 65}$ = ~10^{130}, if it is assumed that one amino acid residue of a protein can occupy 65 different positions including those of a side chain upon wobbling of the immature

protein. Otherwise, the fact, that many EntNew proteins have been created thus far when necessary, cannot be rationally explained.

4. The fourth one is an extraordinary vast undeveloped amino acid sequence space, which is usable to create an EntNew protein. The whole of the amino acid sequence diversity is extraordinary vast as ~10^{130}, as described above. On the other hand, an amino acid sequence diversity of proteins used in extant organisms is estimated as high as ~10^{10} [16]. This means that the only quite small part of the whole sequence diversity has been used by extant organisms, and that therefore it would be easy to find out an an amino acid sequence of an EntNew protein, which is quite different from any other proteins previously used (Table 2).

5. The fifth one is a weak catalytic activity of the immature protein, which leads to evolution to a mature protein.

Table 2. Factors of Immature Protein enabling Creation of Entirely New Protein

* means the power operator.

1. Immature protein synthesis	expression of a GC-NSF(a) under protein 0^{th}-order structure
2. Structure of Immature protein	water-soluble globular structure with some flexibility
3. Characteristics of Immature protein	pluripotency based on flexible structure
4. Undeveloped Sequence Space	extraordinary vast (~20*90 = ~10*127)
5. Maturation of Immature protein	maturation lead by catalytic activity

3.5. The mechanism under which an EntNew protein is actually created in extant organisms

As described repeatedly so far, it is impossible to create any mature protein at one stroke through random process because of the extraordinary large amino acid sequence diversity. In addition, it is also impossible to design amino acid sequence for protein synthesis of such mature proteins in advance. Therefore, it is obvious that any mature protein must be formed through even if not a direct random process but an indirect mechanism containing at least one random process during creation of an EnyNew protein.

Then, how could extant organisms acquire a skillful method creating mature proteins? Regarding this problem, I have proposed GC-NSF(a) hypothesis on the origin of gene (GC-NSF(a) hypothesis), assuming that EntNew genes are created from a GC-NSF(a) or a nonstop frame on antisense strand of a GC-rich gene [2,3].

3.6. Grounds showing that a GC-NSF(a) codes for a random amino acid sequence, which can be folded into a water-soluble globular protein

Then, I introduce the GC-NSF(a) hypothesis explaining the mechanism, with which EntNew proteins have been created under protein 0^{th}-order structure.

1. EntNew proteins must be created through a kind of random process but not an exactly random process.
2. That is, EntNew proteins are created via the respective immature proteins carrying an essentially random amino acid sequence, which are produced by expression of a GC-NSF(a) (Figure 6).

3.6.1. The reason why a GC-NSF(a) encodes one essentially random amino acid sequence

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 March 2021 doi:10.20944/preprints202103.0003.v1

1. An imaginary protein, which is produced from a GC-NSF(a), satisfies the six conditions for water-soluble globular protein formation, which were obtained based on an amino acid composition and secondary or tertiary structure propensities of the respective amino acids [3, 17].
2. Base compositions of GC-NSF(a) at three codon positions are similar to those of SNS, which encodes one of protein 0^{th}-order structures, where S means G or C. This indicates that an amino acid sequence or amino acid composition encoded by a GC-NSF(a) is similar to the SNS-protein 0^{th}-order structure encoded by $(SNS)_n$ sequence, which was used as a genetic sequence under the primitive SNS genetic code [17].
3. Amino acid sequence encoded by GC-NSF(a) is quite different from the amino acid sequence encoded by GC-rich codon sequence on sense strand because of (1) anti-parallel structure of double-stranded DNA, (2) asymmetric structure of the genetic code and (3) degeneracy of the genetic code at the third codon position.

Therefore, the amino acid sequence encoded by a GC-NSF(a) can be regarded as a random one.

3.7. Evidence of the GC-NSF(a) hypothesis

Evidence supporting the GC-NSF(a) hypothesis has been obtained as described below.
1. Direct evidence was obtained, showing that a partial amino acid sequence (63 bases long) encoded by the GC-NSF(a) of *Pseudomonas aeruginosa* transaldolase B (*tal)* gene has a sufficiently high homology with a partial amino acid sequence of cell division protein, FtsZ, encoded by *ftsZ* gene [18]. In addition, another result was also obtained between a part of amino acid sequence encoded by the GC-NSF(a) of major facilitator super family transporter gene and a partial amino acid sequence of extant ABC transporter ATP-binding protein. These mean that EntNew proteins have been generated from the respective immature proteins encoded by GC-NSF(a)s carried by GC-rich genes.
2. It was also confirmed from analyses of appearance frequency of two neighboring amino acids that bacterial proteins are actually formed through random process, as was expected by the GC-NSF(a) hypothesis [3].
3. It was found that frequency of use of hydrophilic amino acids in water-soluble globular proteins are roughly the same irrespective of the number of amino acids used in the extant proteins, indicating that microbial proteins are formed by essentially random joining of amino acids selected out from a specific amino acid composition or a protein 0^{th}-order structure [3].

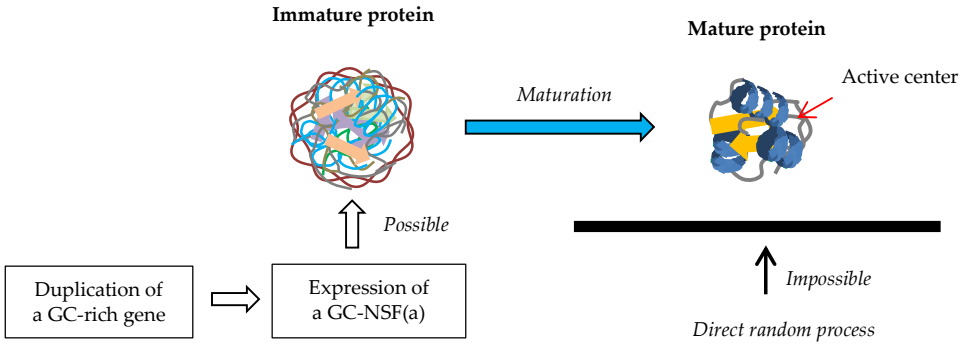

Figure 6. Creation pathway of a mature protein in extant organisms. It is impossible to produce through a direct random process. Therefore, a mature protein is always indirectly created through maturation of an immature protein, which is produced from a GC-NSF(a) of one of duplicated GC-rich genes.

3.8. The specific steps for the creation of EntNew gene/protein in modern organisms

It is indispensable to synthesize first an immature protein through a kind of random process to create an EntNew protein and to evolve the immature protein to mature protein. Next, explain the indirect concrete creation process of a mature protein via an immature protein as follows.

Step 1: *Duplication of a GC-rich gene*: GC-rich gene is first duplicated to prepare an usable GC-NSF(a) as a field for creation of EntNew protein.

Step 2: *Synthesis of an immature protein*: Upon the duplication of one GC-rich gene, an immature but water-soluble globular protein could be produced through expression of the essentially random codon sequence on antisense strand of one of duplicated GC-rich gene or a GC-NSF(a). Hereby, a surplus gene can accept base replacements for the immature protein to evolve to mature protein.

Step 3: *Maturation of the immature protein*: Successively, the immature protein evolves to a mature protein with memorizing ability of the double-stranded DNA for amino acid replacements, as raising the weak catalytic activity of the immature protein.

4. Application of creation mechanism of EntNew protein to a novel strategy of protein engineering

It is expected that it would become possible to create artificially various EntNew proteins with a desired function and to apply the function to a variety of industries if the creation mechanism of EntNew proteins could be freely used, because there is an undeveloped field with great potential to produce artificial proteins. Then, I would like to propose a novel strategy for creation of EntNew artificial proteins in this Section and to open the way of new protein engineering using the strategy. I propose such a novel strategy for development of protein engineering utilizing EntNew artificially created proteins.

I published previously a short paper entitled "Possible Application of EntNew Gene/Protein to Clinical Research" in which application of the creation mechanism of EntNew genes/proteins to clinical research was described [19]. However, my intention of the possibility was not well understood partly because of the short paper. Then, I present again a novel strategy for application of the mechanism for creation of EntNew gene/protein to various industries based on the principle and concrete procedures of protein engineering for artificially created EntNew gene/protein.

4.1. Proposition of a novel strategy and concrete procedures for generating EntNew genes/proteins

Here, I propose methods for using EntNew genes/proteins created artificially.

1. An EntNew protein with a new function, which has been absent in any extant organism, could be acquired by optimization of weak function of an immature protein, which is produced from a GC-NSF(a). Acquired new function itself could be applied to produce new pharmaceuticals, industrial products and so on, of which synthesis has been difficult with methods used previously.
2. Evolutionary process from an immature protein to a mature protein could be also applied to maturation of an immature protein, which was constructed by a new combination between two artificially created EntNew proteins.
3. Construction of new proteins with a new function would become possible through various combinations even between two domains, one is a previously existed domain and the other is an EntNew protein artificially created.
4. The evolutionary process from an immature protein to a mature protein could be applied to an incomplete protein, which was constructed by joining two different previously existed proteins or domains.

4.2. Creation of an EntNew artificial gene/protein

Experiment 1: Establishment of a new procedure through recreation of an EntNew gene/protein with an existing function.

 doi:10.20944/preprints202103.0003.v1

Procedures

Exp. 1-1: A phenotypic revertant with a low catalytic activity is isolated from a lethal deletion mutant carrying AT-rich genome, which was transformed with a GC-rich plasmid to use GC-NSF(a)s. Then, the transformants are grown in a medium containing an organic compound, which is require for the mutant to grow because of the lethal mutation.

Exp. 1-2: It is confirmed whether or not EntNew gene/protein with the same catalytic activity as the enzyme encoded by deleted gene has been recreated in the revertant.

Exp. 1-3: The catalytic activity of the EntNew protein is optimized according to the procedures of evolutionary engineering.

Exp. 1-4: Furthermore, it is confirmed that amino acid sequence of the EntNew protein is quite different from the previously existed protein by analyses of amino acid sequence of the protein and/or base sequence of the corresponding gene. It could be confirmed from the results that the protein isolated is an EntNew one.

Experiment 2: Creation of EntNew protein with a new catalytic function using the knowledge and techniques, which were obtained in Experiment 1.

Procedures

Exp. 1-1: First a new organic compound, for example, as a fluorescent substrate, with which an unprecedented catalytic activity can be detected, is designed. A bacterium carrying AT-rich genome and a GC-rich plasmid is grown with the organic compound as the substrate in a liquid medium.

Exp. 1-2: The bacteria are grown on an agar plate containing the designed organic substrate, if even a low catalytic activity could be detected in the liquid medium.

Exp. 1-3: The catalytic activity of the EntNew protein in the bacterium is optimized according to the procedures of evolutionary engineering.

Exp. 1-4: The EntNew protein with a high catalytic activity is isolated from the bacterium to use as the EntNew catalytic activity for application.

4.3. Uses of an EntNew protein artificially created according to the above procedures

Experiment 1: Direct usage of an EntNew protein with an unprecedented catalytic activity

Procedures

Exp. 1-1: The catalytic activity of a newly created protein is used to synthesize a new organic compound or to degrade an undesired chemical compound.

Experiment 2: Creation of a new artificial proteinaceous catalyst through combination among the EntNew protein artificially created and previously existing proteins or domains.

5. Discussion

Every existing protein exhibits a splendid function so that diverse organisms can live on the present Earth owing to the respective proteins. However, the amino acid sequences, which were and have been used by the past and extant organisms, are limited in a quite small sequence space in comparison to the whole sequence space, as ~10^{130}. Therefore, it should be quite easy to create EntNew artificial proteins. This means that the EntNew artificial proteins should be applied to various industries, if EntNew proteins could be artificially created. Then, I have proposed a novel strategy for application of those artificially created EntNew proteins in this paper.

5.1. Protein 0^{th}-order structure: The first principle for creation of a mature protein

Protein 0^{th}-order structure: A mature protein is always created from an immature protein, which is produced by a random process, such as using an essentially random amino acid sequence encoded by a codon sequence on

GC-NSF(a) or direct joining of [GADV]-amino acids under a protein 0^{th}-order structure. For example,

1. Formation of primeval [GADV]-proteins (actually aggregates of [GADV]-peptides)

The water-soluble globular [GADV]-proteins with some flexibility could be produced by direct random joining of [GADV]-amino acids owing to the protein 0^{th}-order structure, a specific amino acid composition.

2. Creation of an EntNew protein through a kind of random process using protein 0^{th}-order structure on GC-NSF(a)

An EntNew protein was and has been created from an immature protein, which was and has been produced with one amino acid sequence arranged randomly under a protein 0^{th}-order structure with a codon sequence on antisense strand of a GC-rich gene, through maturation process.

3. Creation of an "induced-fit" type enzyme

It is considered that an induced-fit type enzyme is formed by terminating an evolutionary process from an immature protein to a mature protein just before the catalytic site closely fits with the corresponding substrate as a "key-key hole" type enzyme.

4. Formation of a homologous protein

A homologous protein is formed from a mature ancestor protein through a protein immatured partially, upon which some flexibility is acquired by accumulation of amino acid replacements onto the ancestor protein with rigid structure. Homologous proteins with a high catalytic activity have been created by rematuration of the partially immatured proteins.

5. Creation of a mature protein adaptable to a sever environment

In this case too, structure of a mature protein first becomes flexible or is partially immatured upon the change to a sever environment from an ordinary and moderate environment. After that, a new mature protein with rigid structure is reformed to adapt the new environment through an evolutionary process.

As described above, both creation of an EntNew protein and formation of a new protein adaptable to a new environment are always carried out through the following two steps.

1. Formation of a new mature protein always start from an immature protein, which is formed through a kind of random process or a partially immatured process, which is induced by accumulation of amino acid replacements or a change of environment from a moderate to a sever condition.
2. A new mature protein is formed or reformed through an evolutionary process from the respective immature proteins.

Therefore, creation process of an artificial protein are summarized as follows.

1. Production of an immature or incomplete protein under a protein 0^{th}-order structure.
2. Expression of versatile catalytic activities on surface of an immature protein with some flexibility or of pluripotency of an immature protein.
3. A weak but sufficiently high catalytic activity could be found on an immature protein at a high probability owing to the pluripotency of an immature protein.
4. EntNew or new mature proteins could be easily created or formed from an extraordinary vast amino acid sequence space, which is not used by previously existed and any extant proteins.
5. After that, a new mature protein can be obtained through optimization process from the immature protein to a mature as raising the weak catalytic activity.
6. The reason why extant organisms are flourishing on the present Earth is because the first genuine ancestor life invented the amazing mechanism, with which diverse EntNew proteins can be created and new homologous proteins can be also formed with the EntNew proteins using as an ancestore protein after gene duplication.

 doi:10.20944/preprints202103.0003.v1

5.2. Significance of the fact that one small protein or one domain is composed of around 100 amino acids.

1. The number of possible catalytic sites, which appears on surface of a protein composed of 100 amino acids, could reach more than 10^{130} owing to the flexible protein structure or wobbling of surface amino acids (Figure 5).
2. Contribution to optimization of catalytic center upon accumulation of favorable amino acid replacements: The effect of one amino acid replacement can be buffered by the existence of 100 amino acids in a protein (Figure 7 (A)), so that the site exhibiting an only low catalytic activity can be gradually optimized owing to the buffer action (Figure 7 (B)).
3. Furthermore, it is essential for maturation of an immature protein that the protein has a catalytic activity higher than others during accumulation of amino acid replacements onto the evolving protein.
4. Even a low catalytic activity of the immature protein should be sufficient for the immature protein to evolve continuously to a mature protein, because the same catalytic activity cannot be found anywhere on the primitive Earth before appearance of the EntNew protein.
5. Even the low catalytic activity observed at every evolutionary step would be the highest activity on the circumstances (Figure 7 (B)). Therefore, it would be also out of the question, even if the evolutionary process took a quite long time and a large number of generations were required to evolve from the immature protein to a mature protein.

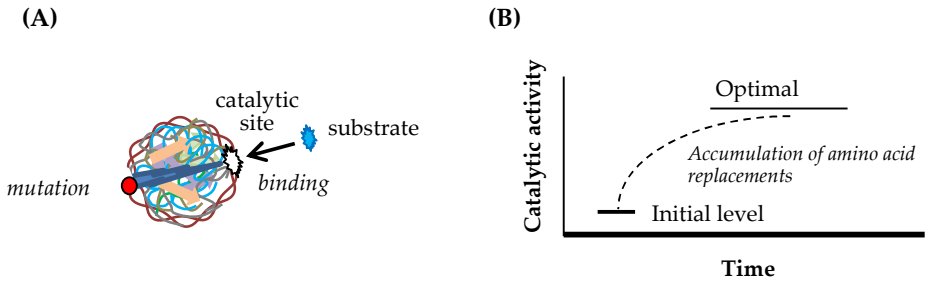

Figure 7. (A) Buffer action of an immature protein composed of about 100 amino acids diminishes efficiently the effect of an amino acid replacement (red circle) to the catalytic site of the protein. **(B)** The buffer action of the protein makes it possible to evolve continuously the immature protein with a low catalytic activity to a mature protein at an optimal level through cumulative subtle adjustments of the catalytic site to substrate.

I retired from my university about 13 years ago. Hence, I have no laboratory and no colleague including a student. However, I have a dream that a new world of protein engineering is opened by young researchers, who want to step into a new research field, as using a new strategy of protein engineering, which is described in this article. I hope for many young researchers to enter the ambitious research field.

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 author declares no conflict of interest.

References

1. Ikehara, K. *Towards revealing the origin of life. -Presenting the GADV hypothesis-;* Springer: London, GB, 2021; *in press*.

2. 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.

3. 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 29, 66-85 (2001)). *J. Biosci*. **2002**, *27*, 165-186.

4. Turanli-Yildiz, B.; Alkim, C.; Cakar, Z.P. *Protein Engineering*. Chapter 2: Protein Engineering Methods and Applications, InTech, Rijeka, Croatia. 2012; doi: 10.5772/27306

5. Sinha, R.; 1, Shukla, P. Current Trends in Protein Engineering: Updates and Progress, *Curr. Protein Pept. Sci.* **2019,** *20*, 398-407. doi: 10.2174/1389203720666181119120120.

6. Wang, Y.; Katyal, P.; Montclare, J.K. Protein-Engineered Functional Materials. *Adv Healthc Mater*. **2019,** *8*, 11. e1801374. doi:10.1002/adhm.201801374.

7. Sun, H.; Li, Y.; Yu, S.; Liu, J. Hierarchical Self-Assembly of Proteins Through Rationally Designed Supramolecular Interfaces. *Front Bioeng Biotechnol*. **2020**, *8*, 295. doi: 10.3389/fbioe.2020.00295.

8. Tanaka, J.; Yanagawa, H.; Doi, N. *Protein Engineering*. Chapter 3: Evolutionary Engineering of Artificial Proteins with Limited Sets of Primitive Amino Acids. InTech, Rijeka, Croatia. 2012; doi: 10.5772/29498

9. Porter, J.L.; Rusli, R.A.; Ollis, D.L. Directed Evolution of Enzymes for Industrial Biocatalysis. *Chembiochem.* **2016,** *217*, 197-203.

10. Crick, F. Central dogma of molecular biology. *Nature* **1970,** *227*, 561-563.

11. Dill, K.A. Dominant forces in protein folding. *Biochemistry* **1990,** *29*, 7133–7155

12. Kunin, V.; Cases, I.; Enright, A.J., de Lorenzo, V.; Ouzounis, C.A. Myriads of protein families, and still counting. *Genome Biol.* **2003,** *4*, 401. doi:10.1186/gb-2003-4-2-401. PMC 151299. PMID 12620116. (protein family)

13. Ikehara, K. Possible steps to the emergence of life: The [GADV]-protein world hypothesis. *Chem. Rec*. **2005,** *5*, 107–118.

14. 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. doi: 10.1007/s11084-014-9384-3

15. Ikehara, K. Protein 0th-Order Structure: The Key for Creating Entirely New Gene/Protein. *Preprints* **2019**, 2019120021 (doi: 10.20944/preprints201912.0021.v1

16. Luisi, P.L. *The Emergence of Life -from chemical origins to synthetic biology-*. 2nd ed.; Cambridge University Press, Cambridge, United Kingdom. 2016; pp. 360-366.

17. 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.

18. Oi, R.; Ikehara, K. Direct evidence for GC-NSF(a) hypothesis on creation of entirely new gene/protein. *Curr. Proteom*. **2018,** *3*, 13-23.

19. Ikehara, K. Possible Application of Entirely New Gene/Protein to Clinical Research. *Clin. Res. Trials* **2018,** *4*, 1-3.