J. Biol. Macromol., **5**(2), 21-30 (2005)

Review

Mechanisms for Creation of "Original Ancestor Genes"

Kenji Ikehara

Department of Chemistry, Faculty of Science, Nara Women's University

Kita-uoya-nishi-machi, Nara, Nara 630-8506, Japan

Received June 24, 2005; Accepted June 29, 2005

Two main theories on creation of new genes

Diversity of organisms on the earth would be tightly coupled to the evolution of new gene functions. Two main routes for creation of new genes have been proposed (Fig. 1). One is the gene duplication theory, which was provided by S. Ohno (1970) (1). The theory predicts that after duplication of a gene, one duplicate may acquire a new adaptive function, while the other duplicate retains the original function (Fig. 1A). The second is the exon shuffling theory proposed by Gilbert *et al*. (1997) (2), assuming that new functional genes are created from exons shuffled among several genes (Fig. 1B).

On the other hand, proteins are generally classified as protein families and superfamilies, according to the sequence-structure information and catalytic properties. As can be seen in Fig. 2, a protein family is composed of proteins with similar amino acid sequences and similar functions, which originated from one original gene, whereas a protein superfamily contains several protein families with similar amino acid sequences but with different catalytic functions.

Therefore, not only proteins belonging to a protein family but also those in a

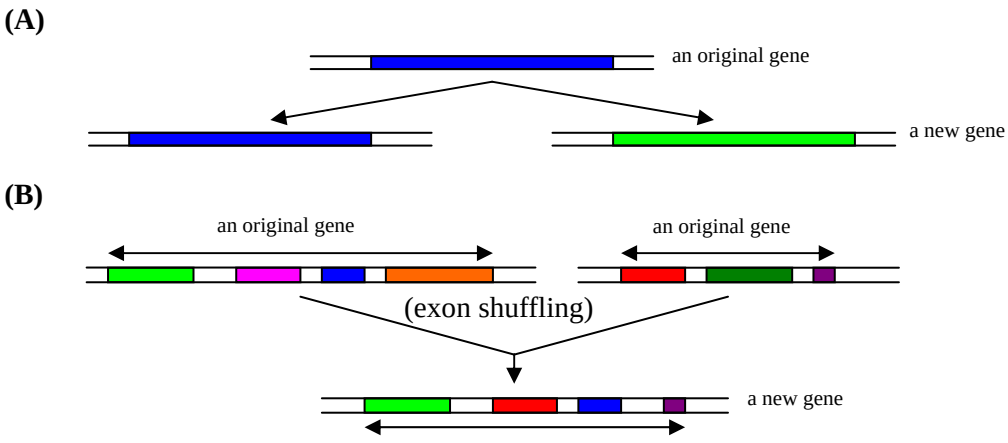

Fig. 1. Two theories proposed for creation of new genes. (A) Gene duplication theory (1). (B) Exon shuffling theory (2).

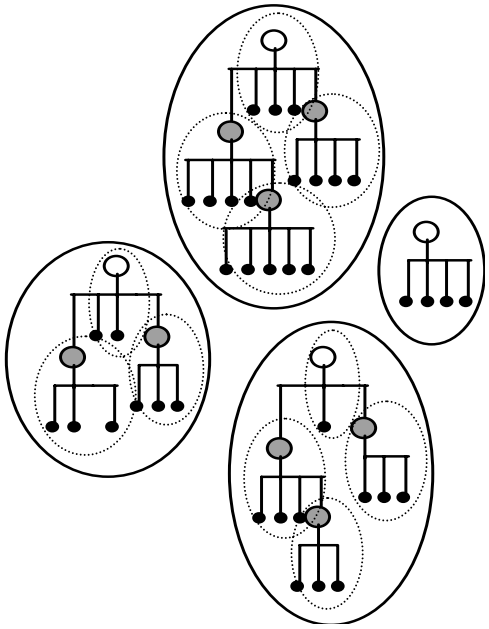

Fig. 2. Individual proteins (small black dots) should be produced from “original ancestor genes” (open circles) and from “original genes” (grey circles). Large ellipsoids small ellipsoids indicate protein superfamilies and protein families, respectively.

superfamily originated from one common original ancestor protein. This indicates that all proteins in the family and the superfamily were produced by expression of genes created by gene duplications (Fig. 2).

Although two theories, gene duplication theory and exon shuffling theory, are essential molecular mechanisms for creation of new genes, both routes require preexisting, original genes. Therefore, both of the two theories do not explain the way how to create original ancestor genes, which are quite different from preexisting genes, since new genes are produced by modification from sense sequences of the parental genes. Here, “original gene” means a parental gene encoding proteins which acquired new function to form a protein family and “original ancestor gene” indicates the first ancestor gene encoding the first protein among a protein superfamily (Fig. 2).

This means that two theories described above does not explain the most fundamental problems on the creation of “original ancestor genes” and that the problems are remained unsolved until now. So, in this review the mechanism for creation of “original ancestor genes”, from which many descendant genes could be produced, is discussed from a standpoint of our GC-NSF(a) and (GNC)n, (SNS)n primitive gene hypotheses (3, 4).

GC-NSF(a) Hypothesis for Creation of “Original Ancestor” Genes under the Universal Genetic Code

We have proposed one hypothesis on the origin of genes about 10 years ago, while performing researches based on the following considerations (3). If new genes were created on the earth at present, from what kind of DNA sequences could “original ancestor genes” have been created under the universal genetic code? Analyses of microbial genes and proteins obtained from the GenomeNet Database, led us the idea that new “original ancestor genes” could be produced from non-stop frames on antisense sequences of microbial GC-rich genes (GC-NSF(a)), as described below (Fig. 3) (3).

The GC contents of microbial genes are distributed over an extremely wide range, from about 25 to 75%. Compositions of

about half the number of amino acids in proteins largely change as the GC contents of the genes increased or decreased under GC- or AT-mutation pressure (4, 5). It is expected that the fundamental properties of globular proteins, which are given by six structural indexes (hydropathy, α-helix, β-sheet, β-turn formabilities, acidic, and basic amino acid contents), are invariable against the differences of GC contents among genes followed by those of amino acid compositions. The reason for this is because proteins must be folded into appropriate three-dimensional structures even in every microbial cell carrying a chromosomal DNA with different GC content. The structural indexes of whole proteins, which determine the secondary and tertiary structures of proteins, were calculated by multiplying amino acid composition with the indexes of 20 amino acids from Stryer's textbook (6).

The analyses of the six indexes of microbial proteins revealed that all indexes are expectedly almost constant against the variation of GC content of genes obtained from seven genomes of bacteria and archaea. This means that when we judge whether or not polypeptide chains can be folded into water-soluble globular structures required for enzymatic functions, the six structural indexes are utilized as necessary conditions. Thus, we first searched for the nucleotide sequences by using the indexes where new "original ancestor genes" could be created under the universal genetic code. For this purpose, we investigated the six structural indexes of hypothetical proteins encoded by possible five reading frames of extant bacterial and archaeal genes, which we obtained from a gene bank. From the results, we found that hypothetical polypeptide chains encoded by antisense sequences on genes with high GC contents (more than 60%) could be folded into globular structures in water or cells at a high probability. Moreover, the probability (pNSF) at which no stop codon appears in a reading frame increases abruptly beyond at about 60% GC content, which is caused by unusually biased base compositions at three codon position (7).

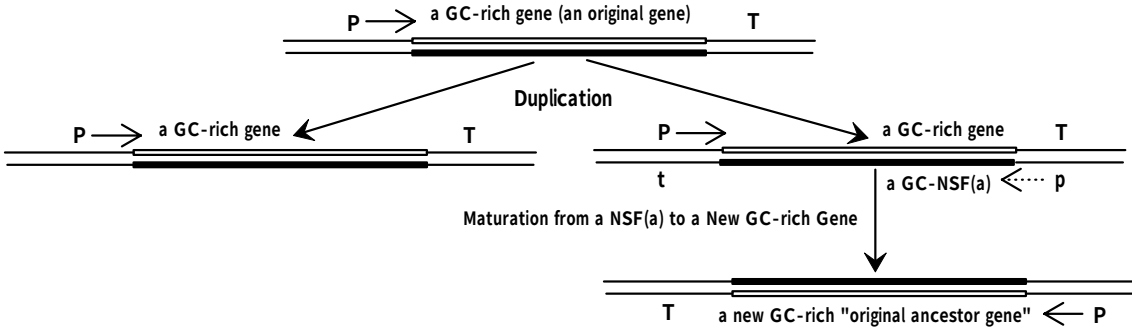

Fig. 3. GC-NSF(a) primitive gene hypothesis for creation of "original ancestor genes" under the universal genetic code. The hypothesis predicts that new "original ancestor genes" originate from nonstop frames on antisense strands of GC-rich genes (GC-NSF(a)s).

It was also found that proteins encoded by GC-NSF(a) hypothetical genes have favorable properties in that they are able to adapt to novel substrates since the proteins would have some flexibility owing to slightly higher glycine contents and smaller hydrophobicity indexes of the proteins than those of actual proteins. Thus, we would like to assert here that the GC-NSF(a)s easily found on GC-rich microbial genomes must be the field, from which new "original ancestor genes" could be produced on the earth at present (Fig. 3) (3, 8).

Hypothesis on the Origin of the Genetic Code

When several months passed after the proposition of the GC-NSF(a) primitive gene hypothesis, I noticed that base compositions of the GC-NSF(a)s can be approximated at the three codon positions as SNS or [(G/C)N(C/G)] at a limit to the GC-rich side (4, 5, 9). To confirm that the SNS code has a powerful ability coding for water-soluble globular proteins, SNS compositions in three codon positions were generated by using computer-generated random numbers. S and N mean either of G or C, and either of four bases, A, U (T), G and C, respectively. We then selected SNS compositions that could satisfy the six structural conditions obtained from extant proteins. The structural indexes of a hypothetical protein were similarly calculated with amino acid composition and the indexes of the corresponding amino acids as described above.

From the results, it was found that the computer-generated SNS code satisfies the six conditions when the contents of G and C at the first codon position were at around 55% and 45%, respectively, and when every four base was contained at a ratio of about one-forth at the second codon position. Base compositions at the third position could not be restricted to a narrow range due to the degeneracy of the genetic code at that position. However, this also means that there was a high probability that polypeptide chains composed of SNS-encoding amino acids ([L], [P], [H], [Q], [R], [V], [A], [D], [E], and [G]) should be folded into globular

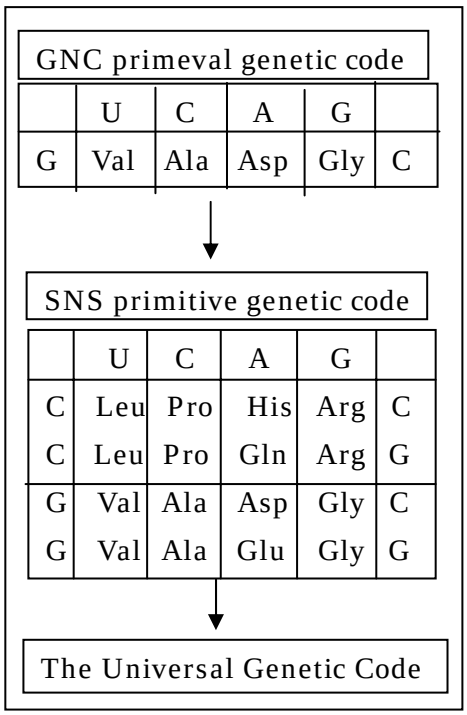

Fig. 4. GNC-SNS primitive genetic code hypothesis, which we have presented to explain the evolutionary pathway to the universal genetic code (5). The hypothesis postulates that the universal genetic code originated from the GNC primeval genetic code through the SNS primitive code.

structures. Thus, we have provided an SNS primitive genetic code hypothesis (9).

In addition, we confirmed that secondary structures and hydropathy profiles of proteins encoded by $(SNS)_n$ hypothetical genes, which were generated by a computer, gave mixed profiles of three secondary structures (α-helix, β-sheet, and β-turn) and of hydrophobic and hydrophilic regions similar to those of existing proteins (data not shown).

Furthermore, we investigated what kind of genetic code that is much simpler than SNS primitive genetic code could encode water-soluble globular proteins. For this purpose, four conditions (hydropathy, α-helix, β-sheet, and β-turn formations) out of the six conditions were used to determine which genetic code could encode water-soluble globular proteins. Consequently, it was found that GNC would be used as the most primeval genetic code on the primitive earth, because four [GADV]-amino acids encoded by the code well satisfied the four structural conditions. This conclusion is the same as the antiquity of the "GNC" code, which was first proposed by Eigen and Schuster (10) from an independent standpoint of the protein coding ability. From the analyses and discussion on the origin of the genetic code, we have proposed the GNC-SNS primitive genetic code hypothesis (Fig. 4) (5).

Under the GNC and SNS genetic codes, group coding for the production of functional proteins should be adopted to avoid meeting stop codons, otherwise non-assigned triplets or the resulting stop codons would often appear in the random sequences at extremely high frequencies. At present, however, it is unknown about what makes it possible to enable group coding.

Establishment of the GNC primeval genetic code

Shimizu (11) has reported that RNAs bearing anticodon nucleotides at the 5' ends and a discriminator base at the 3' ends named as C4N could accept the cognate amino acid by the lock-and-key relationship, and that the C4N complexes were aminoacylated specifically with their cognate amino acids. From theoretical analysis of conformation of the hydrogen-bonded C4N complexes, it is shown that some conformational changes are induced in the anticodon trinucleoside

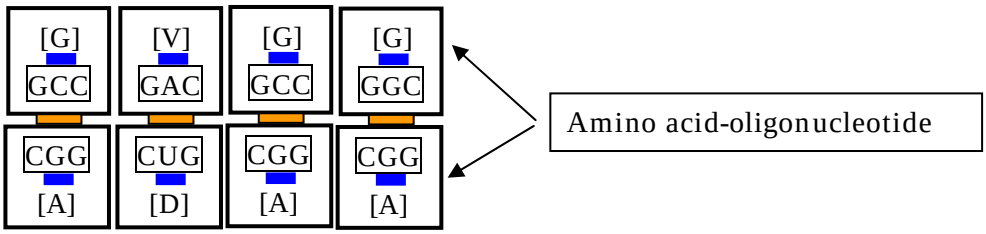

Fig. 5. Establishment of GNC primeval genetic code by specific stereo-chemical binding of [GADV]-amino acids with GNC-containing oligonucleotides. GNC codons arranged randomly but sequentially were used as imitations of genes before creation of the most primitive “original ancestor gene” (single-stranded (GNC)n gene).

diphosphate by the binding of the discriminator base. Conformational changes of C4N and the amino acid are also induced by the binding of the amino acid to C4N (12).

So, it is assumed that [GADV]-proteins accumulated by pseudoreplication of the proteins in [GADV]-protein world synthesized nucleotides and oligonucleotides (4, 8), and that oligonucleotides containing GNC were aminoacylated with cognate [GADV]-amino acid and confronted with complementary complexes similarly as that one is tRNA (anticodon) and the other is mRNA (codon) (Fig. 5). The random lengthwise arrangement of the complexes should accelerate [GADV]-protein synthesis at a higher rate than the synthesis of [GADV]-proteins under the absence of the C4N-like complexes, leading to establishment of the GNC primeval genetic code. It is also supposed that [GADV]-proteins synthesized under the GNC primeval genetic code must contain only [GADV]-amino acids, since the code can translate GNC codons into the four amino acids. The reason why random peptide bond formation among [GADV]-amino acids made it possible to produce water-soluble globular proteins is that the composition composed of [GADV]-amino acids is one of the 0^{th}-order structures of proteins (4).

Deduced Pathway for Creation of "Original Ancestor (GNC)n Genes"

As described above, we have proposed that the first genetic code appeared on the earth must be the GNC primeval genetic code. This implies that even randomly GNC repeating sequences could be utilized as a kind of functional genes encoding water-soluble globular proteins on the

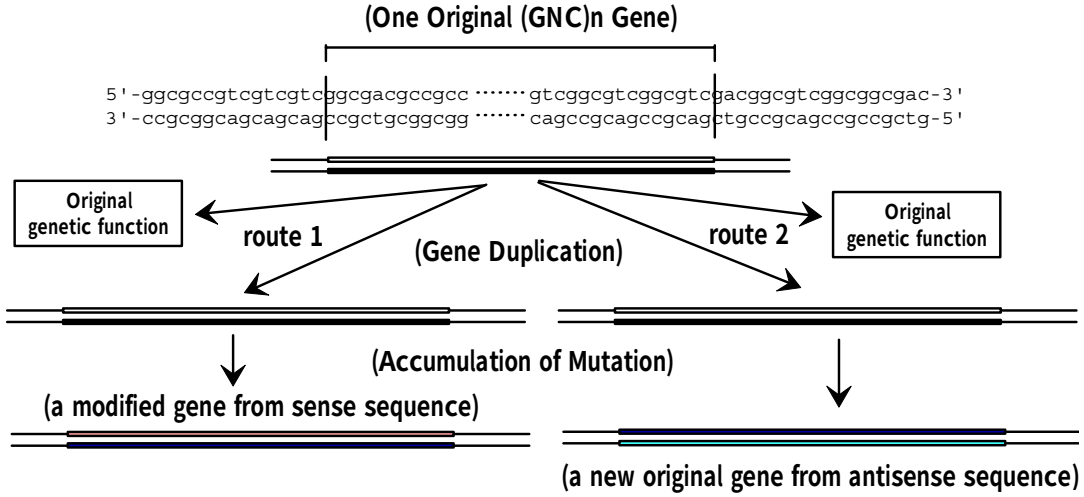

Fig. 6. Two routes for producing new genes. Once one original (GNC)n gene was produced, new genes were easily produced by using the original gene through two routes, one is from sense sequence and the other is from antisense sequence of the original gene. From route 1, new gene could be produced as a modified gene of the original gene, and from route 2, new gene could be created as an "original ancestor gene".

primitive earth. To exhibit genetic functions more effectively, GNC codons must be combined by covalent bonds, resulting in formation of the most primitive single-stranded RNA genes corresponding to modern mRNA, after a lapse of some time from the establishment of the GNC primeval genetic code. Lately, double-stranded (GNC)n genes were formed by binding GNC codons with covalent bonds as the case of formation of the single-stranded (GNC)n genes as described above. Double-stranded (GNC)n genes evolved from the single-stranded GNC sequences were symmetric between codons on sense and antisense sequences.

This enabled for both sequences to use for production of new genes. Namely, two routes for creation new genes are considered (Fig. 6). One route is for producing new genes encoding modified and similar proteins to the original proteins. The other is a route producing new "original ancestor genes" encoding new proteins with quite different amino acid sequences from every sequence of proteins existed before. These suggest that antisense sequences of primeval genes, $(GNC)_n$, would be favorable for synthesis of new "original ancestor [GADV]-proteins" under the most primitive GNC code. Thus, the [GADV]-proteins must lead to create the GNC primeval genetic code followed by (GNC)n primeval genes, resulted in formation of the most primitive fundamental life system.

Deduced Pathway for Creation of "Original Ancestor (SNS)n Genes"

As seen in Fig. 7, it is considered that gene, genetic code and protein were coevolved and must be coevolved in the period from the establishment of the fundamental life system to the present time, since three objects are intimately related to each other. The reasons are as follows. Proteins can not be reproduced without both genes and genetic code, after the establishment of the fundamental life system. Genetic functions on genes can not be expressed without both proteineous enzymes and genetic code. Of

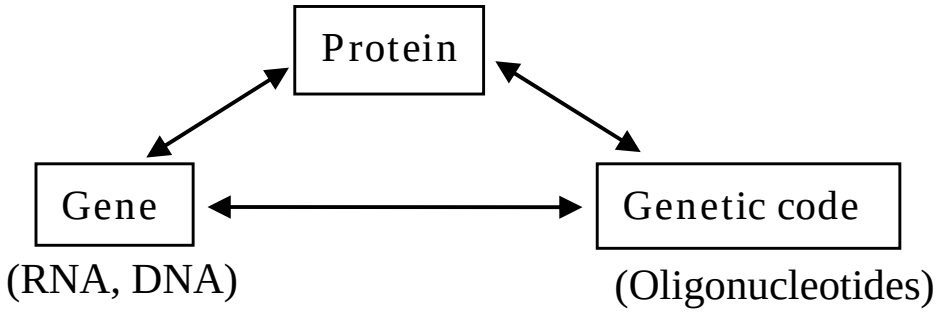

Fig. 7. Coevolution among gene, genetic code and protein. Three objects composed of gene, genetic code and protein have coevolved, after establishment of the most fundamental life system, since they are intimately related each other.

course, genetic codes have the reason for existing, when both genes and proteins are present.

During the coevolution between genetic code and genes, the GNC primeval genetic code gradually developed to SNS primitive genetic code through codon captures, and (GNC)n genes changed their base sequences to (SNS)n sequences. Six pairs of SNS codons out of eight can encode the same amino acids after base exchanges between G and C at the third codon position, since those pairs are degenerated at the position. The degeneracy of the SNS codons enabled to create more easily new "original ancestor genes" from their antisense sequences, as given in Fig. 8, than GNC codons without degeneracy. Under the SNS primitive genetic code, new genes could be created from sense sequences and from antisense sequences through two routes as explained above in the case of (GNC)n genes. In other words, the degeneracy of SNS codons expanded a possibility for creation of "original ancestor genes" from antisense sequences (Fig. 8).

Deduced Pathway for Creation of "Original Ancestor Genes" on the Present Earth

During the coevolution between genetic code and genes, the SNS primitive genetic code gradually developed to the universal genetic code through capturing codons with A and U at the first codon position. In parallel, (SNS)n genes evolved to modern genetic sequences. In this case too, new genes could be created from sense sequences and from antisense sequences through two routes as explained above in the case of (GNC)n and (SNS)n genes. But, in the case of the universal genetic code, it is possible to vary many codons among four bases, A, T, G and C, or among two bases, at the third codon

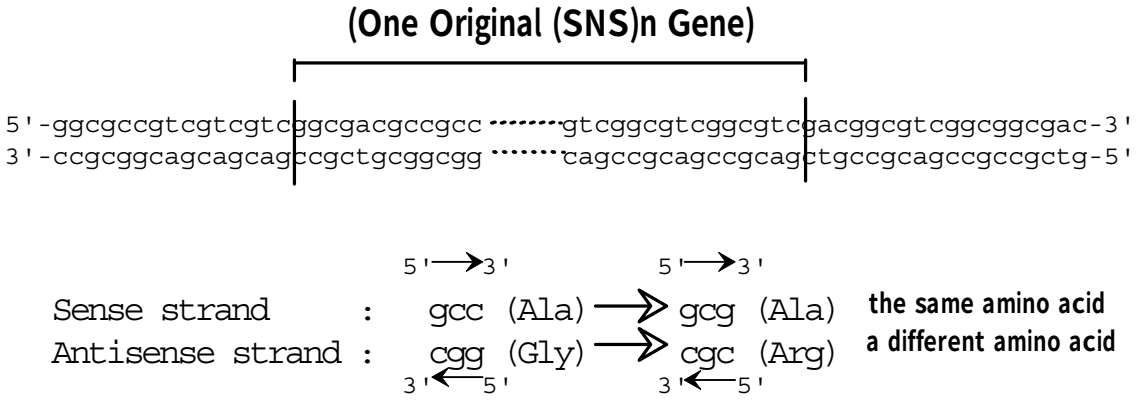

Fig. 8. New gene creation from original (SNS)n gene. New genes were easily produced from original (SNS)n genes through two routes as similarly as those described in Fig. 6, one is from sense sequence and the other is from antisense sequence of the original gene. "Original ancestor genes" could be produced from antisense sequences at a high probability due to the degeneration of SNS code at the third codon position.

position without changing amino acids encoded, resulting in changing bases at the first codon positions in codons on antisense strand and in amino acid replacements in the range restricted in modified SNS 0^{th}-order structure of proteins.

The degeneracy of codons in the universal genetic code enabled to create new "original ancestor genes" from their antisense sequences much more easily than before.

Discussion

In this review, I surveyed possible pathways, through which new "original ancestor genes" were created. Before the most primitive (GNC)n genes were created, random lengthy arrangements of GNC triplets or GNC codons were used in place of the primitive genes. At a next step, single-stranded (GNC)n genes were created by joining of GNC triplets with phosphodiester bonds (Fig. 5). Then, the single stranded (GNC)n genes evolved to double-stranded (GNC)n genes by joining of GNC triplets on the other side with covalent bonds. After the formation of genes on double-stranded RNA or DNA, new genes could be created through two routes, one is from sense strands for producing new genes encoding modified and similar proteins to the original proteins, and the other is from antisense sequences of the original genes for producing new "original ancestor genes", as given in Fig. 6. The first route from the sense sequences and the second route from the antisense sequences contributed to expansion of protein families and to the creation of new protein families, respectively. It can be concluded that, genes in lives on this planet have been diversified by applying the two routes for producing new genes.

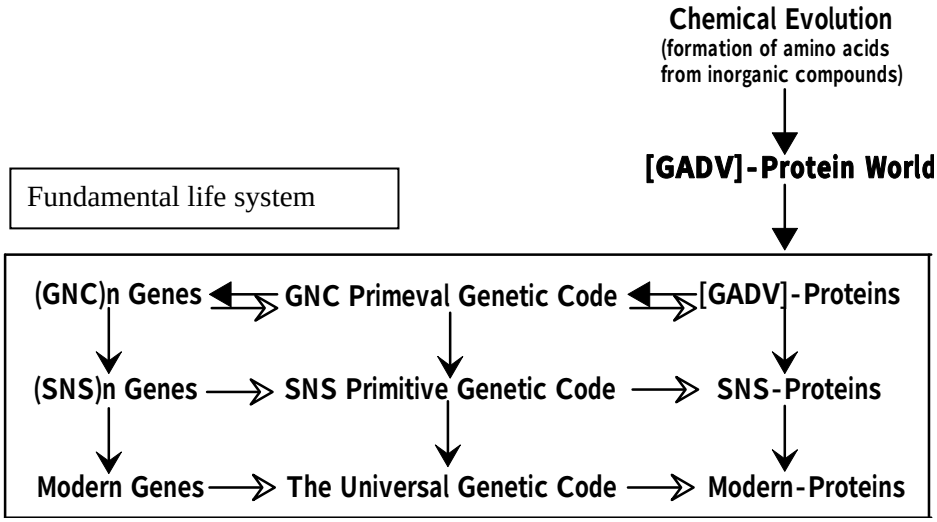

Fig. 9. Possible evolutionary pathway from chemical evolution to the fundamental life system composed of genes, genetic code and proteins. [GADV]-proteins accumulated by pseudoreplication in [GADV]-protein world synthesized nucleotides and established GNC primeval genetic code followed by formation of (GNC)n genes. The most primitive fundamental life system coevolved to the modern life system. Bold arrows, thin arrows and white arrows indicate formation processes, evolutionary pathways and flows of genetic information, respectively.

Through coevolution of genetic code and genes, (GNC)n sequences were evolved to (SNS)n genes and modern genes used under the universal genetic code. Possible evolutionary pathway from chemical evolution to establishment of the fundamental life system through [GADV]-protein world are shown in Fig. 9.

References

1. Ohno, S., Evolution by gene duplication; *Springer Heiderberg* (1970).
2. Gilbert, W., de Souza, S. J., and Long, M., Origins of genes; *Proc. Natl. Acad. Sci., USA* **94** 7698-7703 (1997).
3. Ikehara, K.; Amada, F.; Yoshida, S.; Mikata, Y.; and Tanaka, A. A possible origin of newly-born bacterial genes: significance of GC-rich nonstop frame on antisense strand. *Nucl. Acids Res.*, **24**, 4249-4255 (1996).
4. 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.*, **27**, 165-186 (2002).
5. Ikehara, K.; Omori, Y.; Arai, R.; and Hirose, A., A novel theory on the origin of the genetic code: a GNC-SNS hypothesis; *J. Mol. Evol*. **54**, 530-538 (2002).
6. Stryer, L., Biochemistry 3rd Ed.; New York, *W.H.Freeman and Company* (1988).
7. Ikehara, K.; and Okazawa, E., Unusually biased nucleotide sequences on sense strands of *Flavobacterium sp*. genes produce nonstop frames on the corresponding antisense strands; *Nucl. Acids Res.* **21** 2193-2199 (1993).
8. Ikehara, K., Possible steps to the emergence of life: The [GADV]-protein world hypothesis. *Chem. Rec.*, **5**, 107-118 (2005).
9. Ikehara, K., and Yoshida, S., SNS hypothesis on the origin of the genetic code; *Viva Origino* **26** 301-310 (1988).
10. Eigen, M., and Schuster, P., The hypercycle. A principle of natural self-organization. Part C: the realistic hypercycle. *Naturwissenschaften*, **65**, 341–369. (1978).
11. Shimizu, M., Specific aminoacylation of C4N hairpin RNAs with the cognate aminoacyl-adenylates in the presence of a dipeptide: origin of the genetic code. *J. Biochem*., (Tokyo). **117**, 23-26 (1995).
12. Yoneda, S., Shimizu, M., Go, N., Fujii, S., Uchida, M., Miura, K., and Watanabe, K., Theoretical and experimental approach to recognition of amino acid by tRNA. *Nucleic Acids Symp Ser*. **12**, 145-148 (1983).

Communicated by Hiroshi Ueno