Review

Origins of gene, genetic code, protein and life: comprehensive view of life systems from a GNC-SNS primitive genetic code hypothesis*

K IKEHARA

Department of Chemistry, Faculty of Science, Nara Women's University, Kita-uoya-nishi-machi, Nara, Nara 630-8506, Japan

(Fax, 742-20-3402; Email, ikehara@cc.nara-wu.ac.jp)

We have investigated the origin of genes, the genetic code, proteins and life using six indices (hydropathy, *a*-helix, *b*-sheet and *b*-turn formabilities, acidic amino acid content and basic amino acid content) necessary for appropriate three-dimensional structure formation of globular proteins. From the analysis of microbial genes, we have concluded that newly-born genes are products of nonstop frames (NSF) on antisense strands of microbial GC-rich genes [GC-NSF(a)] and from SNS repeating sequences [$(SNS)_n$] similar to the GC-NSF(a) (S and N mean G or C and either of four bases, respectively). We have also proposed that the universal genetic code used by most organisms on the earth presently could be derived from a GNC-SNS primitive genetic code. We have further presented the [GADV]-protein world hypothesis of the origin of life as well as a hypothesis of protein production, suggesting that proteins were originally produced by random peptide formation of amino acids restricted in specific amino acid compositions termed as GNC-, SNS- and GC-NSF(a)-0th order structures of proteins. The [GADV]-protein world hypothesis is primarily derived from the GNC-primitive genetic code hypothesis. It is also expected that basic properties of extant genes and proteins could be revealed by considerations based on the scenario with four stages.

[Ikehara K 2002 Origins of gene, genetic code, protein and life: comprehensive view of life systems from a GNC-SNS primitive genetic code hypothesis; *J. Biosci*. **27** 165–186]

1. Introduction

Since the Institute for Genomic Research (TIGR) published the first complete microbial genomic sequence of *Haemophilus influenzae* in 1995, more than 40 microbial genomes have been sequenced and published. In recent years, genomic sequences of *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Homo sapiens* were also determined and published. In parallel with the determination of genomic sequences, information of primary sequences or amino acid sequences of proteins has been rapidly accumulated. Tertiary structures of many proteins including complex ribosomal particles have also been determined. Thirty years ago, when I was a university student, these advances could not even be imagined. In those days, even sequences of several bases on DNA could not be determined, since no restriction enzyme was discovered. At that time, the research to determine tertiary structures of proteins such as lysozyme which had only 129 amino acids had just begun. In spite of the rapid progress of biological knowledge, it is not well known how genes and proteins were created, and how base compositions in the codon positions and average amino acid compositions in proteins have

Keywords. GNC-SNS primitive genetic code hypothesis; origin of genes; origin of genetic code; origin of life; origin of proteins

*This review is a modified English version of the paper, which was written in Japanese and published in *Viva Origino* 2001 **29** 66–85.

been determined. That is, it is little known how the fundamental system of life was created.

On the other hand, we started research work on the origin of genes from following considerations about 10 years ago. (i) In which kind of a field, were genes produced on ancient earth? (ii) At what kind of a field on the present earth, is favourable for new genes to arise, if they arise at all? Next, we carried out an analysis on the origin of the genetic codes. Consequently, we have reached a novel GNC-SNS genetic code hypothesis. It anticipates a possible evolutionary pathway, suggesting that the universal genetic code has originated from GNC code through SNS code, where S and N mean G or C and either of four bases (A, G, T(U) and C) respectively (Ikehara 1998a,b, 1999; Ikehara and Yoshida 1998). Parallelly, we have arrived at another hypothesis, suggesting a route how ancestor proteins were created on primitive earth and the ancestral proteins have evolved to proteins used in organisms on the present earth (Ikehara, Nakanishi and Katayama, unpublished data). Based on our hypotheses on the origins of genes, genetic code and proteins, we have also presented a novel hypothesis, which could reasonably explain the origin of life (Ikehara 1999, 2000).

In this review, at first, we will discuss the problems of the origins of genes, the genetic code and proteins, along with a flow of genetic expression. After that, we will discuss on the origin of life, which is also involved in the fundamental life systems and is interrelated to the three origins. Then we will compare our hypotheses with those provided by other researchers. The comparison will help the readers of this review to understand our novel hypotheses on the four origins. Our idea, which could give systematic understanding about the fundamental life system mainly based on the GNC-SNS primitive genetic code hypothesis, are quite different from ideas provided by other researchers, in which the four origins have been treated as independent problems. At the last section in this review, we will also provide some evidences supporting our hypotheses.

2. Origin of genes

2.1 *Our hypothesis on the origin of genes*

2.1a *GC-NSF(a) hypothesis on creation of genes on the present earth*: GC contents of microbial genes are distributed in an extremely wide range from about 20 to 75% (figure 1). It is generally considered that the wide distribution of GC content observed in microbial genes would be produced by a GC-mutation pressure and by an AT-mutation pressure, which acts on the genes in directions to increase and to decrease the GC content respectively (Sueoka 1988). About half the amino acid compositions largely vary, as GC content of a gene varies under the mutation pressures (figure 2). But it will be difficult to vary the indices, which determine abilities of secondary and tertiary structure formation in proteins. Thus basic properties of globular proteins, which are given by six indices (hydrophobicity/hydrophilicity, *a*-helix, *b*-sheet, *b*-turn formabilities, acidic and basic amino acid contents), should be invariable against changes of GC content of a gene and amino acid composition of a protein. From analyses of the six indices required to form tertiary structure, it was found that all six indices are actually almost constant with change in GC content of

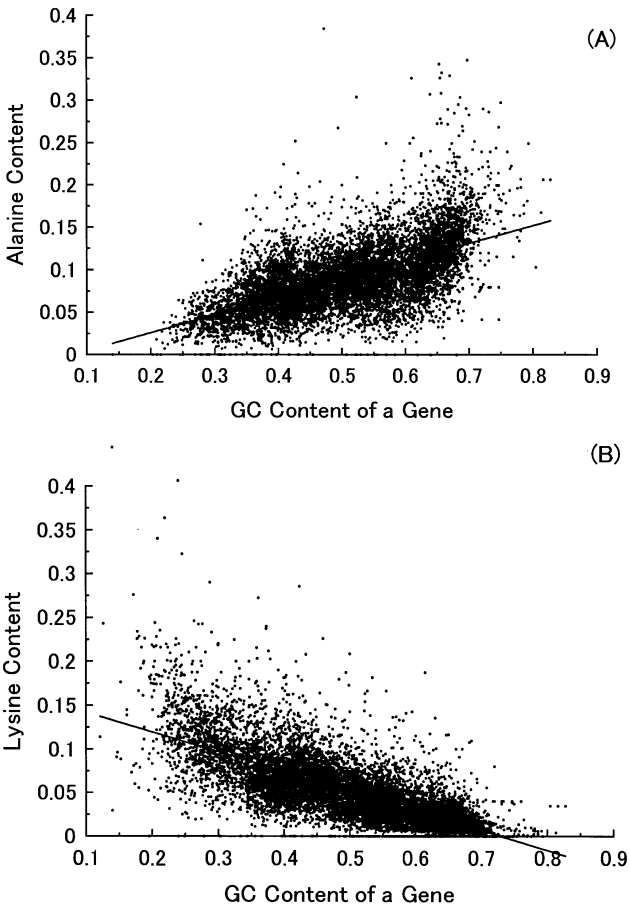

Figure 1. Dependence of alanine (**A**) and lysine (**B**) contents on GC content of a gene. The amino acid contents were analysed using 7 microbial (5 eubacteria and 2 archaea) genome data [*Mycobacterium tuberculosis* (65·6), *Aeropyrum pernix* (56·3), *Escherichia coli* (50·8), *Bacillus subtilis* (43·5), *Haemophilus influenzae* (38·1), *Methanococcus genitarium* (31·3) and *Borrelia burgdorferi* (28·2); the numbers in parentheses indicate the average GC contents of the microbial genomes]. Dependence of other 18 amino acid contents on GC content of a gene were obtained by the similar procedures described in this figure (data not shown).

genes encoded by 7 genomes in bacteria and archea (figure 3). Results on the dependencies of acidic amino acid content and of basic amino acid content on GC content are omitted on account of limited space in the review. This means that the six structural indices of proteins can be used as necessary conditions, when we determine whether the polypeptide chains can be folded into water-soluble globular structures to exhibit enzymatic functions (table 1). Thus, we first searched for a field by using the conditions, where new genes could be created on the present earth. For this purpose, we investigated the six structural indices of hypothetical proteins encoded by 5 possible reading frames of extant bacterial and archeal genes which we got from the data-bank. The indices of amino acids for hydropathy and for secondary structure formations were obtained from Stryer's textbook (Stryer 1988). From the results, it was found that the hypothetical polypeptide chains encoded by antisense sequences on genes with high GC contents (more than 50%) would be folded into similar globular structures to actual proteins at a high probability (figure 4). Moreover, the probability (pNSF), of any stop codon not appearing in the frame, abruptly increased in a region higher than 60% GC content, caused by unusually biased base compositions at three codon positions (figure 5) (Ikehara and Okazawa 1993).

Hereafter, we call a nonstop frame (NSF) on antisense strand of GC-rich gene as GC-NSF(a). Proteins encoded by the GC-NSF(a) have favourable properties to be able to adapt to novel substrates, since the proteins would have some flexibility. That is because glycine contents and hydrophobicity indices of the proteins are slightly larger and smaller than those of actual proteins respectively (Ikehara *et al* 1996). Therefore, we consider that the GC-NSF(a) easily found on GC-rich microbial genomes must be the field, where new genes could be produced on the present earth.

Table 1. Average values (A.V.) and standard deviations (SD) of structure indices (A) and acidic and basic amino acid contents (B) for water-soluble globular proteins encoded by 7 microbial genomes.

(A)

Index	A.V.	S.D.
Hydrophobicity	– 1·513	0·382
a-Helix	1·027	0·034
b-Sheet	1·004	0·023
b-Turn	0·964	0·045

(B)

Content	A.V.	S.D.
Acidic amino acids	0·12	0·032
Basic amino acids	0·141	0·031

2.1b *$(SNS)_n$ hypothesis on the formation of genes on primitive earth*: As described above, it is considered that the GC-NSF(a) would be used as a field for creation of new genes (Ikehara and Okazawa 1993; Ikehara *et al* 1996). Base compositions of the hypothetical GC-NSF(a) genes as well as GC-rich genes are similar to S, N and S at the first, second and third codon positions, respectively (figure 6). Thus, base compositions of the GC-NSF(a) can be approximated as SNS or [(G/C)N(C/G)] at a limit to GC-rich side. This implies that at least bacteria and archea with GC-rich chromosomes use an approximated form of SNS-repeating sequences (figure 6), and that SNS-repeating sequences themselves could be utilized as functional genes encoding globular proteins at a high probability.

SNS compositions in the codon were generated by using random numbers by a computer, to confirm whether main chains of hypothetical proteins encoded by $(SNS)_n$ can be folded into similar structures to actual proteins. Then, we selected out SNS compositions, which can satisfy the six structural conditions obtained by the extant proteins. From the results, it was found that the contents of G and C at the first codon position are optimal at around 55% and 45%. Moreover, it was optimal

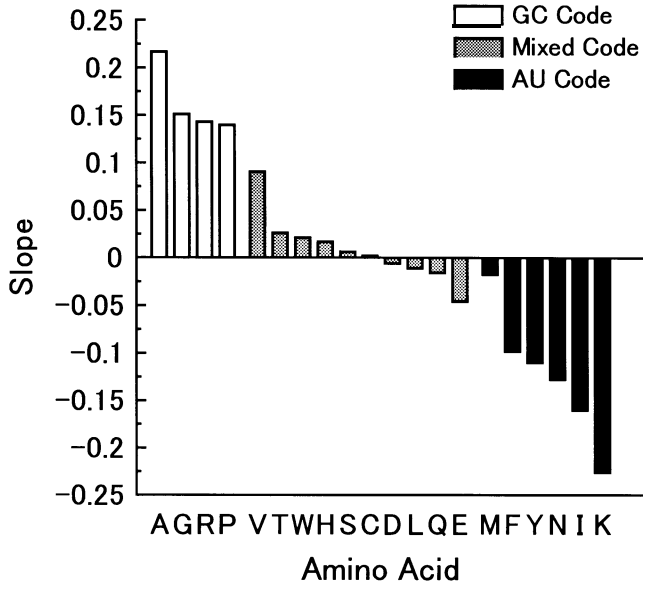

Figure 2. The slope of an approximated linear line obtained from the plot of amino acid content against GC content of a gene as seen in figure 1. Large characters described on the abscissa mean amino acids represented by one-letter symbols. Open bars, shaded bars and closed bars indicate amino acids encoded by GC codes, mixed codes and AU codes, respectively.

for formation of globular proteins, when about 1/4 of every base was contained at the second codon position (figure 7). Base compositions at the third position need not be restricted in a narrow range, due to degeneracy of the genetic code at the position. But, this means that proteins encoded by hypothetical $(SNS)_n$ genes can satisfy the six conditions for protein structure formation, and that polypeptide chains composed of 10 SNS-encoding amino acids (L, P, H, Q, R, V, A, D, E, G) could be folded into globular structures similar to extant proteins at a high probability, when base compositions were given as 55%G and 45%C at the first, 25%A, 25%T, 25%G and 25%C at the second, and 55%C and 45%G at the third codon positions respectively (Ikehara 1998a; Ikehara and Yoshida 1998). Further, we examined secondary structure and hydrophobicity profiles of the proteins encoded by the $(SNS)_n$ hypothetical genes, which are generated by a computer. The results gave an appropriately mixed profile of three secondary structures (*a*-helix, *b*-sheet and *b*-turn) as in the cases of actual proteins (figure 8). Both hydrophobic and hydrophilic regions were also appropriately mixed up in hydrophobicity/hydrophilicity profiles, likewise extant proteins (data not shown). These indicate that SNS repeating sequences, $(SNS)_n$, could be used as the origin of genes on primitive earth.

2.2 *Theories on the origin of genes provided by other researchers*

2.2a *Gene duplication theory*: Consider that a gene in an organism was duplicated by any reason. The organism can maintain life activities by using one genetic function in the same way as it could be before the gene duplication,

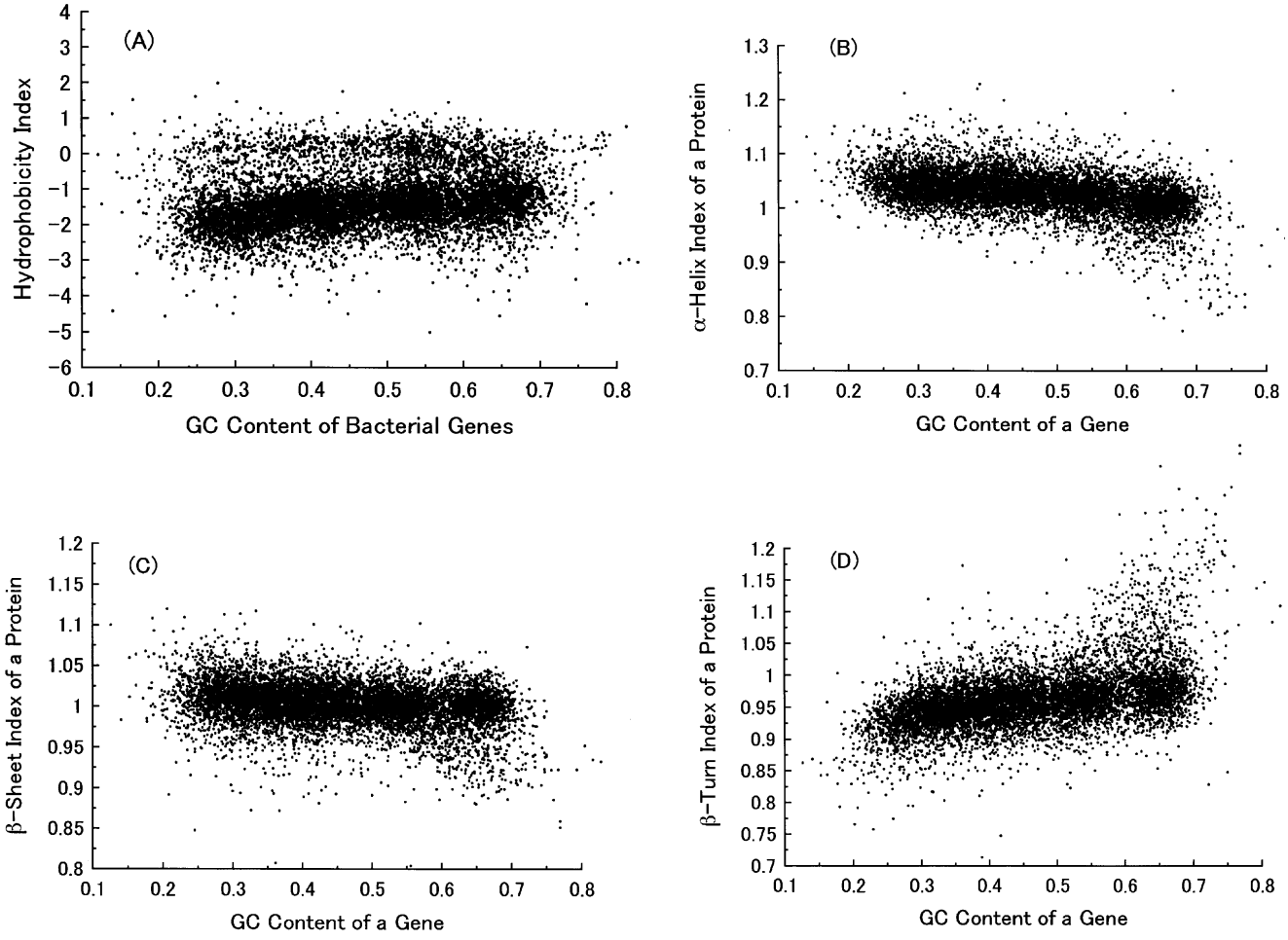

Figure 3. Dependence of hydropathy (**A**), *a*-helix (**B**), *b*-sheet (**C**) and *b*-turn (**D**) formability indices upon GC content of a gene. Lower thick and upper faint bands seen in (**A**) represent hydropathy indices of globular proteins and of membrane proteins, respectively. (**B**), (**C**) and (**D**) show the data of globular proteins only. Similar results were obtained by analyses of acidic and basic amino acid contents (data not shown).

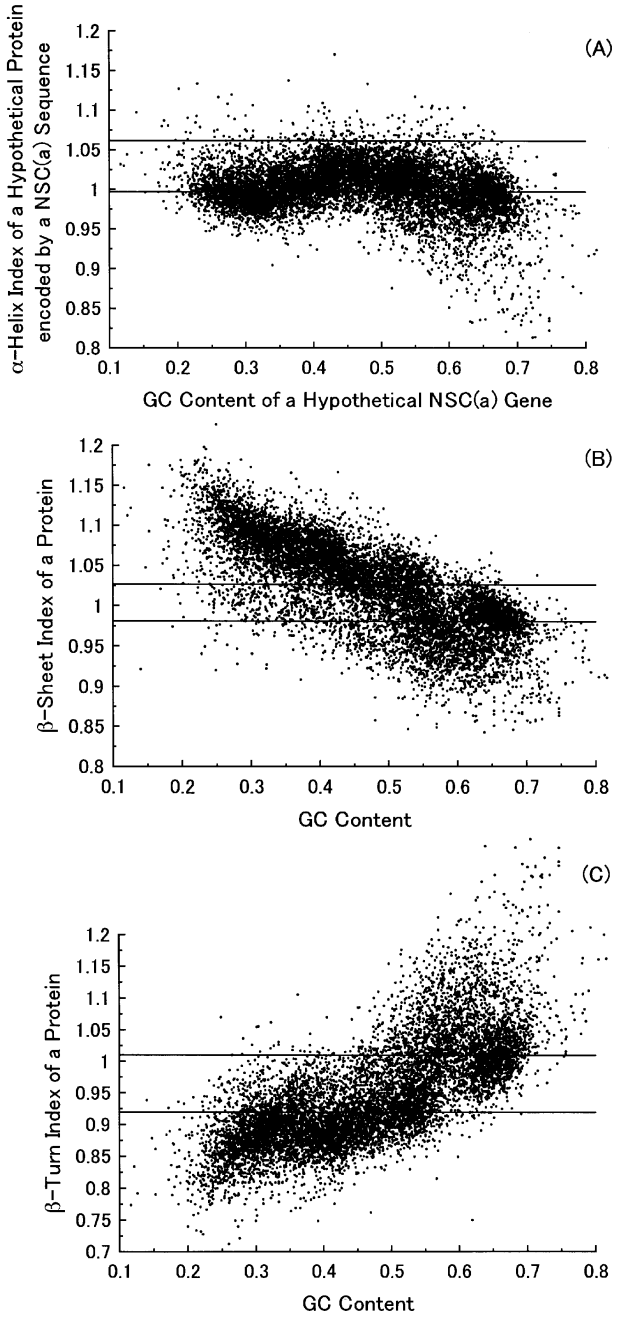

Figure 4. Dependence of secondary structure formability indices [*a*-helix (**A**), *b*-sheet (**B**) and *b*-turn (**C**)] upon GC content of a hypothetical gene on an antisense sequence. Total indices, which were calculated from structure index of each amino acid and the respective amino acid compositions of a hypothetical protein encoded by an antisense sequence of a gene from 7 microbial genome data, are plotted against GC content of the hypothetical gene. Two horizontal lines in the figure show the positions of the average value (AV) plus and minus the standard deviation (SD) of actual proteins. The average values and the standard deviations are given in table 1. It can be seen that over about 50% GC content, the index values enter the regions of AV plus and minus of SD of globular proteins.

even if mutations were accumulated on the other gene. This means that it is possible to accumulate forbidden mutations on one of the duplicated genes. Based on this consideration, the gene duplication theory was provided by S Ohno, which suggests that novel genes with new functions can be created by accumulation of mutations on one of the duplicated genes (figure 9) (Ohno 1970). It is well known that there exist a large number of homologous protein families, in which proteins with different amino acid sequences have similar catalytic functions, and proteins with similar amino acid sequences exhibit different catalytic functions. When amino acid sequences belonging to the same protein families are compared with each other, a significant ratio (more than 30%) of amino acids on a sequence is consistent with each other. That is, there are a large number of genes, which can be considered as derivatives produced from the same ancestor gene by gene duplication. Thus,

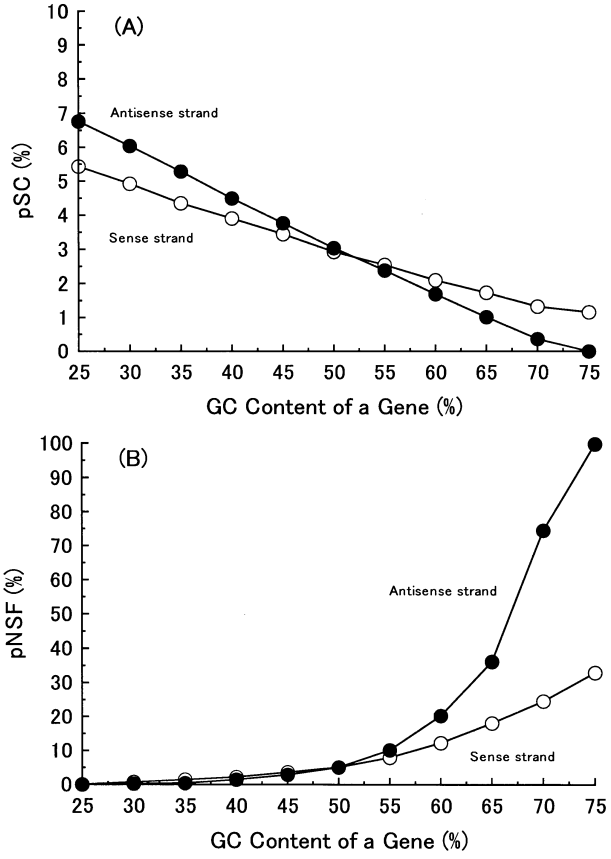

Figure 5. Dependence of probabilities of stop codon appearance (pSC) (**A**), and of nonstop frame appearance (pNSF) (**B**) on sense strand (open circles) and on antisense strand (closed circles) upon GC content of a gene. The pNSF values, which are composed of 100 codons, were obtained by calculation with the pSC values on the sense strand (open circles) and antisense strand (closed circles), which were estimated from base compositions at three codon positions.

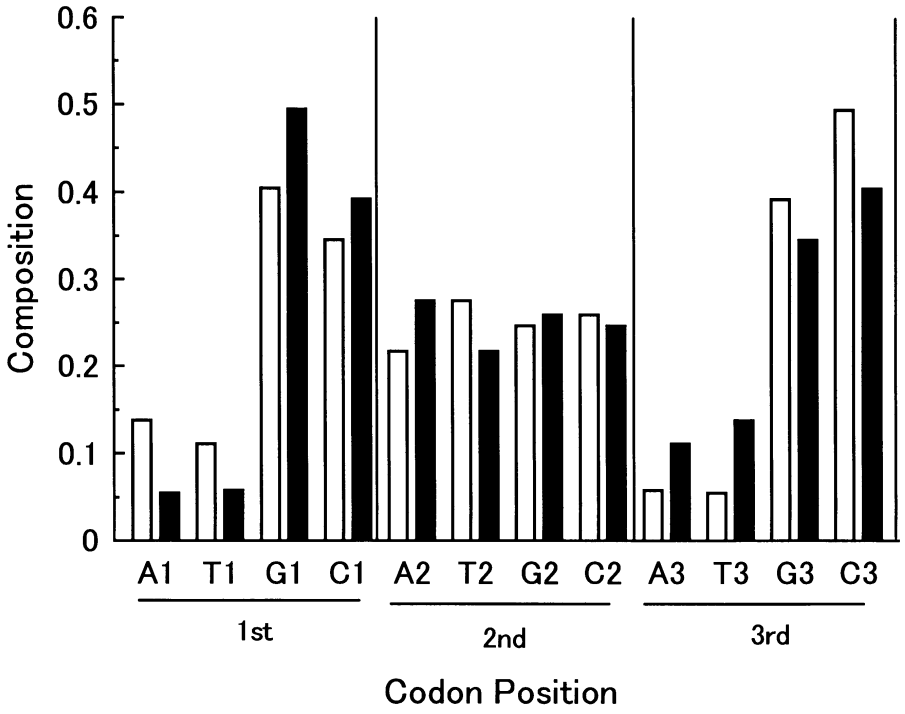

Figure 6. Average base compositions at three base positions in the codon of *Pseudomonas aeruginosa* genes higher than 70% GC content (open bars), and of the corresponding hypothetical genes (nonstop frames) on antisense strands of the GC-rich genes [GC-NSF(a)] (closed bars).

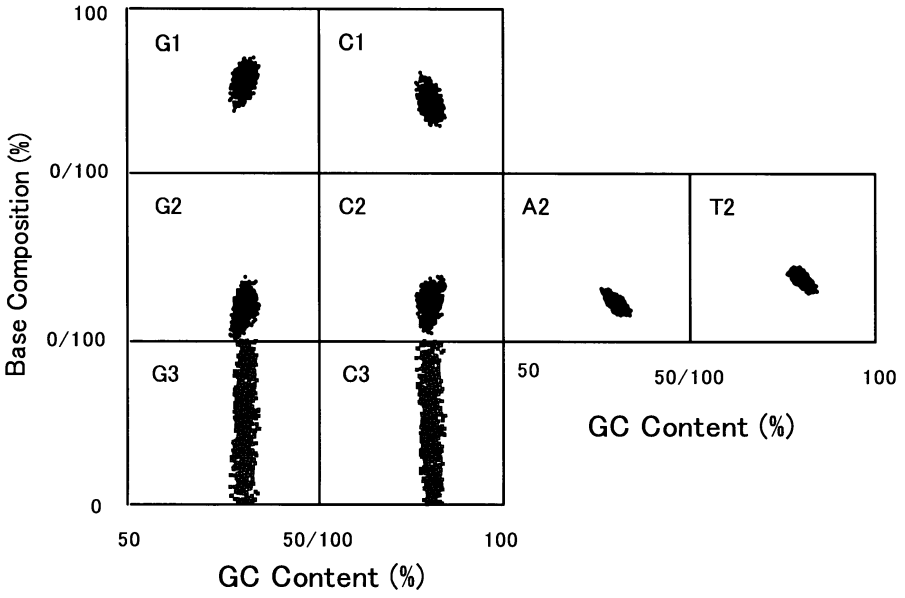

Figure 7. Dot representation of computer-generated base compositions at three base positions in the codon, which were selected by determining whether imaginary proteins translated under the SNS coding system (S and N mean G or C and either of four bases, respectively) satisfy the six structural conditions (hydropathy, *a*-helix, *b*-sheet and *b*-turn formabitities, acidic and basic amino acid contents) for appropriate three-dimensional structure formation. Base compositions at the three codon positions were plotted against GC content of the selected genes out of the computer-generated hypothetical genes.

nowadays, the gene duplication theory is regarded as a correct theory.

2.2b *Exon-shuffling theory*: The exon-shuffling theory, which is quite different from the gene duplication theory described above, has been also presented to explain a process for formation of a new gene (Gilbert *et al* 1997). The theory attaches importance to existence of introns on eukaryotic genes. It states that the original genes were exons encoding small polypeptides composed of only 15–20 amino acids, and that various kinds of genes could arise by shuffling exons and introns (figure 10). Many

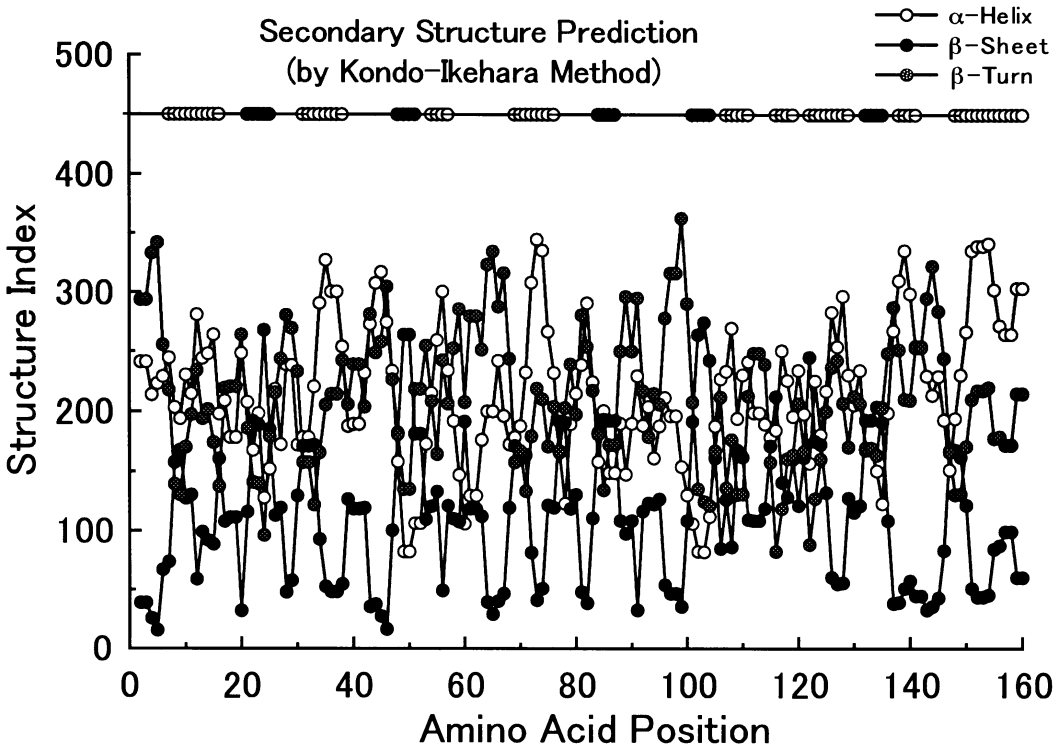

Figure 8. Secondary structure profile of a hypothetical protein encoded by a SNS repeating sequence [$(SNS)_n$] deduced from secondary structural indices with the Kondo-Ikehara method (unpublished method). Open circles, closed circles and shaded circles in the lower graph indicate *a*-helix, *b*-sheet and *b*-turn regions of the hypothetical protein, respectively. Thin lines in the upper graph represent the predicted *b*-turn or coil regions of the hypothetical protein.

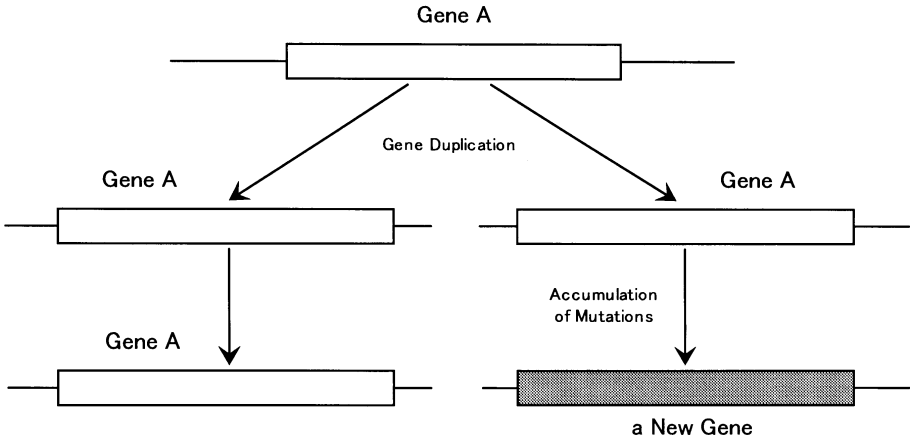

Figure 9. The gene duplication theory on the origin of genes. In this model, it is supposed that a newly-born gene is produced from one of duplicated genes after accumulation of base substitutions, while the other gene retains the original genetic function necessary to live on.

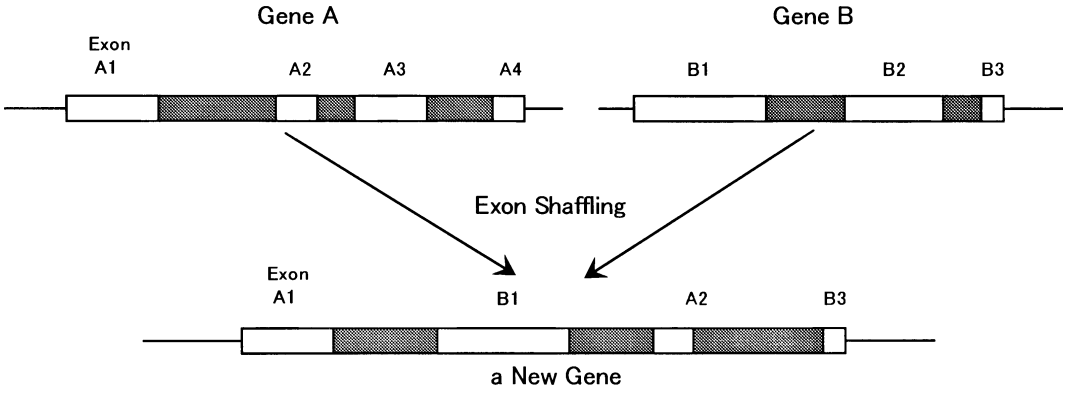

Figure 10. The exon(-shuffling) theory on the origin of genes. In this model, it is considered that a newly-born gene is produced by shuffling exons on plural number of genes after duplication of the genes required to shuffle them.

researchers have discussed on the validity of the exon-shuffling theory.

The above theories have been discussed only on processes for new gene formation from previously existing genes. Therefore, the theories can never explain the process, how original genes would be created. Thus, we cannot call them as a rightful theory on the origin of genes.

3. Origin of the genetic code

3.1 *Our hypothesis on the origin of the genetic code*

3.1a *SNS primitive genetic code hypothesis*: As described above, we have proposed $(SNS)_n$ sequences as a field for creation of genes in the days, when life was appearing on primitive earth. Since the genes coded for only a restricted number of amino acids (10 kinds of amino acids) at that time, the primitive genetic code was enough to encode these amino acids. Based on such an intimate relationship between genes and the genetic code, we have provided the SNS-primitive genetic code hypothesis for the origin of the genetic code (figure 11) (Ikehara 1998a,b; Ikehara and Yoshida 1998).

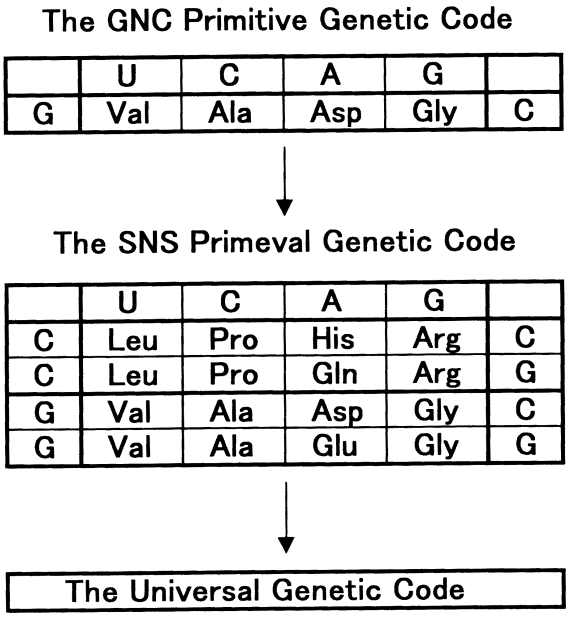

Figure 11. The GNC-SNS primitive genetic code hypothesis. This shows an evolutionary pathway from GNC code (4 codon system) to the universal genetic code (64 codon system), through SNS code composed of 16 codons and 10 amino acids.

3.1b *GNC primeval genetic code hypothesis*: The SNS primitive genetic code is considerably complex, since the code is composed of 10 amino acids encoded by 16 codons (figure 11). Thus, it must have been very difficult to create the SNS code at one stroke at the beginning on the primitive earth. To solve this problem, we searched for a simpler code than the SNS code, which can still encode water-soluble globular proteins with appropriate three-dimensional structures at a high probability. For this purpose, four conditions (hydropathy, *a*-helix, *b*-sheet and *b*-turn formations) out of the six conditions for globular protein formation were used to determine which codes could encode primitive water-soluble globular proteins. Then, we excluded both the acidic amino acid content and basic amino acid content from the conditions, because it would be difficult to contain both acidic amino acid(s) and basic amino acid(s) in smaller kinds of amino acids than 10. Moreover, in case of proteins containing acidic amino acids but not basic amino acids, cations with positive charge such as metal ions, could compen-

sate for the negative charges on the proteins. In the cases of proteins, which contain basic amino acids but not acidic amino acids, anions with negative charge, such as halogens, could neutralize the positive charges on the proteins. From the search for a simpler genetic code, it was found that four amino acids (Gly: [G], Ala: [A], Asp: [D], Val: [V]) encoded by G-start codons, GNC, and its modified form, GNG code, well satisfied the four structural conditions (figures 12 and 13) (Ikehara 1999; Ikehara *et al* 2002). This means that 4 amino acids encoded by GNC have abilities necessary to form secondary and tertiary structures similar to presently existing proteins, and that primitive water-soluble globular proteins could be produced from [GADV]-proteins at a high probability. We have also confirmed that every combination of 4 amino acids encoded by codons standing in rows (CNG, CNC, ANG, ANC, UNG, UNC) and in columns (NUC, NUG, NCC, NCG, NAC, NAG, NGC, NGG) in the universal genetic code table, cannot satisfy the four conditions, except for the GNC and GNG codes (figure 13) (Ikehara *et al* 2002).

Three out of the 4 amino acids have excellent abilities required for the formation of respective secondary structures ([A]: *a*-helix, [V]: *b*-sheet, [G]: *b*-turn or coil). Moreover, [D] has a functional group (carboxylic group), which is indispensable to construct a catalytic centre on primeval proteins (table 2). Both hydrophobic amino acid ([V]) and hydrophilic amino acid ([D]) are also fortunately included in the four amino acids encoded by the GNC code (table 2), which are necessary for folding of polypeptide chains into stable globular structures in water. Furthermore, we confirmed that the [GADV]-amino acids is the simplest set, which is required for the formation of appropriate secondary and tertiary structures

Table 2. Properties of [GADV]-amino acids encoded by GNC codons necessary for globular structure formation and the presence of a functional group.

Amino acid	GNC code	Sec. structure	Hydropathy	Functional group
Val	GUC	*b*-Sheet	Hydrophobic	–
Ala	GCC	*a*-Helix	–	–
Asp	GAC	(*b*-Turn)	Hydrophilic	Carboxyl group
Gly	GGC	*b*-Turn		–

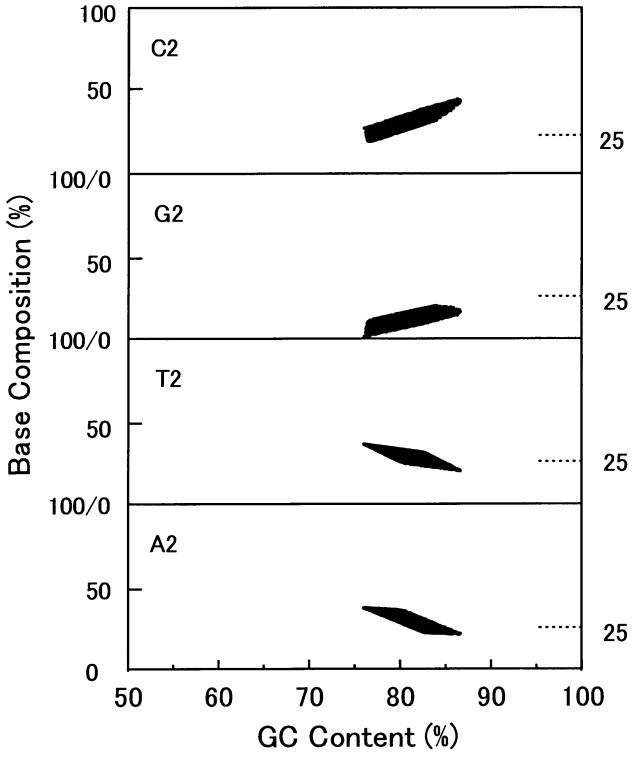

Figure 12. Dot representation of computer-generated base compositions at the second base position in the codon, which were selected by determining whether imaginary proteins translated under the GNC coding system satisfy the four structural conditions (hydropathy, *a*-helix, *b*-sheet and *b*-turn formabilities) for appropriate three-dimensional structure formation. Base compositions at the second codon position were plotted against GC content of the selected gene out of the computer-generated hypothetical genes.

	U	C	A	G		
U	Phe	Ser	Tyr	Cys	U	
U	Phe	Ser	Tyr	Cys	C	[FSYC] X
U	Leu	Ser	term	term	A	
U	Leu	Ser	term	Trp	G	[LSYW] X
C	Leu	Pro	His	Arg	U	
C	Leu	Pro	His	Arg	C	[LPHR] X
C	Leu	Pro	Gln	Arg	A	
C	Leu	Pro	Gln	Arg	G	[LPQR] X
A	Ile	Thr	Asn	Ser	U	
A	Ile	Thr	Asn	Ser	C	[ITNS] X
A	Ile	Thr	Lys	Arg	A	
A	Met	Thr	Lys	Arg	G	[MTKR] X
G	Val	Ala	Asp	Gly	U	
G	Val	Ala	Asp	Gly	C	[VADG] O
G	Val	Ala	Glu	Gly	A	
G	Val	Ala	Glu	Gly	G	[VAEG] O

[FLIV] X [SPTA] X [YHND] X [CRSG] X
[FLMV] X [SPTA] X [YQKE] X [WRSG] X

Figure 13. Structure formability of hypothetical proteins composed of 4 amino acids, which are represented by one-letter symbols in brackets. They were picked up from columns and rows of the universal genetic code table. Symbols of O and X outside of brackets represent that hypothetical proteins composed of the 4 amino acids satisfied the 4 conditions (hydropathy index, *a*-helix, *b*-sheet and *b*-turn formabilities) necessary for formation of globular structures or not, respectively.

(figure 14) (Ikehara 1999; Ikehara *et al* 2002). It is also known from Miller's electric discharge experiments that the four amino acids ([G, A, D and V]) could be easily synthesized on primitive earth (Miller and Orgel 1975). We further investigated on whether there are three amino acid systems, which can satisfy the four structural conditions of globular protein formation. Although three sets of three amino acid systems ([D], Leu and Tyr; [D], Tyr and Met; Glu, Pro and Ile) satisfied the above four conditions, amino acids contained in the sets are scattered in the universal genetic code table. In addition, it can be seen that the structure of at least one amino acid in the three systems is far more complex than those encoded by the GNC code. These results also suggest that the GNC primitive genetic code was used before the SNS primitive genetic code, and that a genetic code simpler than the GNC code had never existed. Thus, the universal genetic code was originated not from a three-amino acid system but from a four-amino acid system, the GNC code encoding [GADV]-proteins. These considerations led us to conclude that the 4 amino acids encoded by the GNC code was used to create globular proteins on primitive earth, and the GNC code was the earliest genetic code appeared on the earth (Ikehara *et al* 2002).

Amino Acid replaced	[GAD] (-Val)	[GAV] (-Asp)	[GDV] (-Ala)	[ADV] (-Gly)
-	-	-	-	-
Gly (1)	(-)	(-)	(-)	65
Ala (4)	(-)	(-)	65	(-)
Ser (5)	-	-	-	57
Cys (5)	-	-	-	-
Asp (7)	(-)	65	(-)	(-)
Asn (8)	-	2	-	171
Thr (8)	-	-	-	110
Pro (9*)	-	-	-	20
Val (10)	65	(-)	(-)	(-)
Glu (10)	-	19	-	-
Gln (11)	-	48	45	-
His (11)	107	85	2	-
Met (11)	-	-	35	-
Leu (13)	-	-	26	-

Figure 14. Structure formability of hypothetical proteins composed of 4 amino acids. The numbers represent the degree of structure formability of the hypothetical proteins, when one of four [GADV]-amino acids is replaced with another amino acid. Minus symbols mean that the 4 amino acids did not satisfy the 4 structural conditions. At the left column, amino acids are placed in order number of atoms on a side chain of an amino acid, or in order structure simplicity of the amino acid. Asterisk indicates the number of atoms existing between the *a*-carbon atom and imino nitrogen atom of proline.

3.2 *Hypotheses on the origin of the genetic code provided by other researchers*

3.2a *Mitochondrial-type primitive genetic code hypothesis*: The mitochondrial-type primitive genetic code hypothesis composed of 20 amino acids and 64 codons has been also provided, chiefly based on a simplicity of the code more than the universal genetic code and on utilization of a minimal set of tRNAs required for translation of genetic information in mitochondria (Osawa 1995).

About 60 codons and 20 amino acids are necessary to establish the primitive genetic code in the beginning of the evolutionary process, if the mitochondrial-type genetic code was used as the most primitive genetic code. It would be too complex to create the genetic code at one stroke in the beginning (figure 15).

The Genetic Code used in Mitochondria (Animals)

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
U	Phe	Ser	Tyr	Cys	C
U	Leu	Ser	term	Trp	A
U	Leu	Ser	term	Trp	G
C	Leu	Pro	His	Arg	U
C	Leu	Pro	His	Arg	C
C	Leu	Pro	Gln	Arg	A
C	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
A	Ile	Thr	Asn	Ser	C
A	Met	Thr	Lys	Arg	A
A	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
G	Val	Ala	Asp	Gly	C
G	Val	Ala	Glu	Gly	A
G	Val	Ala	Glu	Gly	G

Figure 15. The genetic code table used in mitochondria of most animal cells. In the table, UGA and AUA are used as tryptophan and methionine codons in addition to usual UGG tryptophan and AUG methionine codons, respectively. Usually, a minimal set of tRNAs is used to translate all codons in the mitochondria.

3.2b *WWW primitive genetic code hypothesis*: The WWW primitive genetic code hypothesis has been presented (where W means A or U), considering metabolic pathways for synthesis of nucleotides and of RNA-world hypothesis on the origin of life (we will more precisely discuss on it at a later section in this review). The hypothesis states that nucleotides, A and U were produced at an earlier time than the other nucleotides, G and C, on primitive earth. It can be reasonably deduced from the facts that the structures of A and U are simpler than the other nucleotides, G and C, and these nucleotides are synthesized at an earlier stages of metabolic pathways than the latter nucleotides. Thus, the WWW hypothesis insists that the original genetic code must be triplets composed of only A and U, or WWW (figure 16) (Jimenez-Sanches 1995; Voet *et al* 1999).

According to this hypothesis, the most primitive amino acids, which were used for protein synthesis should be the following 6 or 7 amino acids, Phe, Leu, Tyr, Ile, Asn and Lys, or plus Met (figure 16). But, structures of these amino acids are far more complex than [GADV]-amino acids and the WWW code composed of 6 or 7 amino acids cannot code for any acidic amino acids. Moreover, we know that the WWW code cannot satisfy the four structural conditions, because too many hydrophobic amino acids [Phe, Leu, Ile (and Met)] are contained, whereas only one weak *b*-turn formation amino acid (Asn) is included in the WWW code. Therefore, there exists a big defect that polypeptide chains synthesized according to the WWW code could not form water-soluble globular proteins.

4. Origin of proteins

Next, consider a way how proteins were created. It is well known that an amino acid sequence is synthesized according to a genetic information or a gene, which is given as repeats of triplets (three nucleotide sequences). Tertiary structures of proteins are formed based on the amino acid sequences specified by the genetic code. Therefore, the origins of genes and the genetic code should be intimately related to the origin of proteins. From these considerations, it is expected that the origin of proteins or mechanisms producing proteins is made clear based on our hypotheses on the origins of genes and the genetic code. Here, at first, we introduce the hypotheses provided by other researchers on the origin of proteins and their weak points, in order to make it easy to understand our novel hypothesis on the origin of proteins.

WWW Primitive Genetic code

	U	A	
U	Phe	Tyr	U
U	Leu	Term	A
A	Ile	Asn	U
A	Ile(Met)	Lys	A

Figure 16. The WWW primitive genetic code hypothesis deduced from the nucleotide metabolism and the RNA world hypothesis on the origin of life. In the hypothesis, it is considered that adenine and uracil nucleotides would be used for the primitive genetic code, before appearance of guanine and cytosine nucleotides on primitive earth.

4.1 *Hypotheses on creation of proteins by other researchers*

4.1a *Amino acid sequence hypothesis*: The sequence hypothesis on the origin of proteins or a mechanism producing proteins is based on the fact that a unique sequence must be synthesized to produce a specified protein with a globular structure. In other words, synthesis of an appropriate amino acid sequence is necessary to produce an active protein. According to the idea, a unique sequence exhibiting a specified function must be selected out from all possible amino acid sequences produced when amino acids were one-dimensionally and randomly located (figure 17) (Dill 1990).

However, even in small proteins with 100 amino acid residues, there are enormously large sequence diversity as $20^{100} = \sim 10^{130}$, since natural proteins are composed of

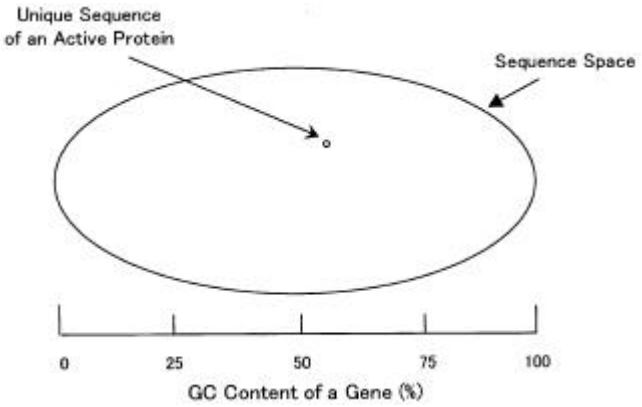

Figure 17. Amino acid sequence hypothesis on the origin of proteins. For a protein composed of 100 amino acid residues, there are 20^{100} = ca. 10^{130} different sequences. In the hypothesis, it is assumed that extant proteins (dots) have been produced through selection of one unique sequence from the enormously large number of sequences. But, it is generally considered that it would be impossible to select the sequence from the sequence space (open ellipsoid).

20 kinds of amino acids (Dill 1990). The number is much larger than the number of all atoms in the universe, which is estimated at 9×10^{78} (Voet *et al* 1999). There is a big defect in the sequence hypothesis that it is impossible to select out a unique sequence from a sequence space with the enormously large diversity. That is because it is impossible to examine all sequences in the space one by one, which amino acid sequence is useful for a required protein. Therefore, another hypothesis has been provided to avoid the difficulty, how one specific amino acid sequence was selected out as its own sequence of an active protein, as described below.

4.1b *Protein structure hypothesis*: This hypothesis gives importance to the tertiary structure of a protein, rather than to the primary structure or the amino acid sequence itself, for the formation of an active protein. It is considered in the hypothesis, that a protein should exhibit the same enzymatic activity as one specific protein when it forms a similar or homologous tertiary structure to the protein. According to the hypothesis, it can be considered that a ratio of hydrophobic to hydrophilic amino acids and the number of interactions in a protein are important to fold the main strand into a similar water-soluble globular structure to the protein with a specific amino acid sequence. A calculation was carried out with ribonuclease as an example based on a lattice model of proteins with a value of 10^{120} to the sequences which have similar polypeptide conformations as the protein (figure 18) (Dill 1990). Therefore, it is considered that an active enzyme accidentally uses one amino acid sequence in a sequence space within the large diversity. The structure hypothesis on the origin of proteins seems to be a correct idea, because a large number of homologous proteins with the same catalytic activity as or with similar tertiary structure to one unique enzyme can be found out in presently existing proteins.

However, the presence of peptide bonds in actual proteins would also be important, since free rotations are restricted around the peptide bonds to make planes in the NH–CO bonds and the planarity of peptide bond largely contributes to the formation of secondary structures. In addition to the hydrophobic and hydrophilic interactions, secondary structure formabilities, such as *a*-helix, *b*-sheet and *b*-turn, should be actually important to form tertiary structures from primary structures of proteins. Therefore, it seems to me that the structure hypothesis on the folding problem of proteins is oversimplified, because only hydrophobic and hydrophilic interactions are considered as the dominant folding force and the lattice model ignores the presence of peptide bonds in proteins. In addition, amino acid sequences of proteins with the same catalytic activity are not always necessary to be similar each other, according to the hypothesis. But it is inconsistent with the fact that about 30 to 40% of amino acid sequences are usually conserved among homologous proteins with the same activity and with similar tertiary structures. The existence of conserved regions among homologous proteins clearly indicates that the proteins were produced from one common ancestor protein, but not selected out independently from the sequence space with the large diversity, as expected from the structure hypothesis. From such considerations, we would like to insist that it is impossible to explain the origin of proteins from a stand point of the structure hypothesis.

4.2 *Our hypothesis on the origin of proteins*

Then, how should we consider about a field and an evolutionary pathway for production of proteins? As a matter of course, the process producing proteins would be intimately related to the origins of genes and the genetic code. Therefore, we consider that proteins were not created as suggested by the amino acid sequence hypothesis or by the protein structure hypothesis, but originated from the proteins which were created in a field producing globular proteins at an extremely high probability (figure 19A). The high probability is assured by small kinds of amino acids and amino acid compositions, which are restricted by the GNC and SNS primitive genetic codes. For example, a diversity of proteins with 100 amino acid residues is calculated as 4^{100} = about 10^{60} in the case of the GNC code, and as about 10^{100} in the case of the SNS code. The diversities are extremely smaller than that (about 10^{130}) of extant proteins com-

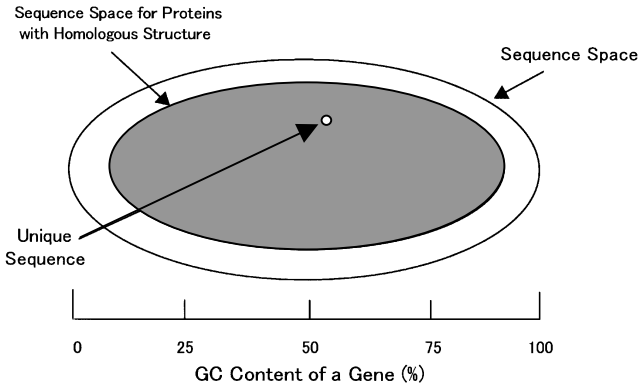

Figure 18. Protein structure hypothesis. To avoid the difficulty of the sequence hypothesis described in figure 17, another hypothesis is presented to explain the origin of proteins. The probability of drawing any sequence, which will fold to a specified structure is fairly large. For an example of ribonuclease, it is estimated to be about 120 orders of magnitude (shaded ellipsoid) by lattice simulations. Therefore, a particular protein uses one sequence out of the enormously large number of sequences in the shaded region by chance.

posed of 20 kinds of amino acids. The former corresponds to about 10^{-70} and the latter is about 10^{-30}. Thus, we would like to give names to the novel hypotheses on the origin of proteins – as the GNC 0th-order structure hypothesis and the SNS 0th-order structure hypothesis, because the GNC- and the SNS-primitive genetic codes should be responsible for the production of primitive proteins in the respective ages. It may not be easy to understand the meaning of the 0th-order structure of proteins, since the term is not familiar to the readers. This will be explained more concretely by using the figure 19B.

The principle is widely accepted that a protein conformation is specified only by its amino acid sequence, which is determined by one-dimensional genetic information. Therefore, it has been generally considered that amino acid sequences or primary structures must be start points to produce globular proteins. But during the process, considering the origins of genes and the genetic code, it was noticed that specific amino acid compositions are far more important to create new proteins than the primary structures. According to our idea, water-soluble globular proteins could be produced at a high probability, even by random joining of amino acids determined by specified amino acid compositions. To explain such a novel idea straightforwardly, I would like to use the word, 0th-order structure, since an amino acid composition is a concept at a lower level than the

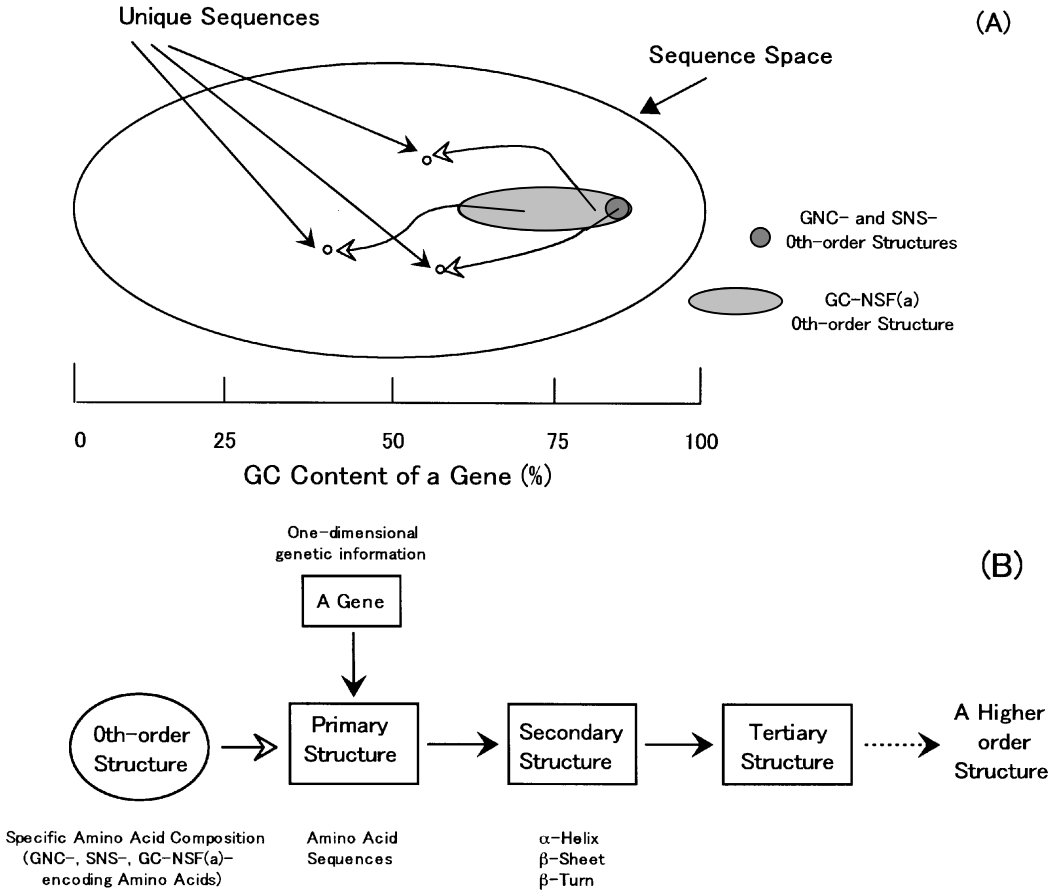

Figure 19. GNC-, SNS-, and GC-NSF(a) 0th-order structure hypotheses on the origin of proteins. (**A**) According to the 0th-order hypothesis, extant proteins should originated from a small number of random sequences (dark circles) specified by particular amino acid compositions, which are determined by the GNC- and SNS primitive genetic codes. At present, newly-born proteins could be derived from proteins encoded by ancestral GC-NSF(a) genes (shaded ellipsoid), a modified form of SNS repeating sequences, $(SNS)_n$. (**B**) It is generally considered that a tertiary structure or a higher-order structure of a protein is determined by a primary structure, or an amino acid sequence, which is determined by one-dimensional genetic information on DNA. Contrary to that, we have provided another hypothesis on formation of globular proteins, where primary structures are formed by random joining of the amino acids restricted by GNC-, SNS 0th-order structures and its modified form of GC-NSF(a) 0th-order structure (**A**). First genes determining amino acid sequences (the primary structure) of newly-born proteins had been realized by the 0th-order structures through creative processes of newly-born genes.

primary structure or an amino acid sequence. We consider the compositions consisting of 4 kinds of amino acids encoded by the GNC code and of 10 kinds amino acids encoded by the SNS code, as the fundamental 0th-order structures. And, we have given to the idea, the terms of the GNC 0th-order hypothesis and the SNS 0th-order hypothesis. In other words, the GNC 0th-order structure and the SNS 0th-order structure were utilized in the former and in the latter ages to create new proteins efficiently on primitive earth, respectively. Further, we consider that the amino acid compositions specified by GC-NSF(a) (GC-NSF(a) 0th-order structure or a modified form of the SNS 0th-order structure) must be used to produce new proteins on the present earth, if necessary (figure 19).

As a matter of course, genes, the genetic code and proteins are the most fundamental and important to biological functions of life. As described above, the GNC code should be the genetic code used in the most primitive life on the earth (figure 11). In addition, the simplest set of 4 kinds amino acids, which can produce functional globular proteins at a high probability, is [GADV]-amino acids encoded by the GNC code (figure 13). Taking these into consideration, we have also provided a novel hypothesis on the origin of life.

5. Origin of life

5.1 *Our hypothesis on the origin of life*

5.1a *[GADV]-protein world hypothesis*: According to the GNC primitive genetic code hypothesis, the most primitive proteins should be composed of only four [GADV]-amino acids. In spite of the simple amino acid compositions, several important properties exist in the [GADV]-proteins, which are required to exhibit enzymatic functions, as described below.

(i) The [GADV]-proteins can form water-soluble globular structures at a high probability, judging from that the proteins have similar indices of hydropathy and secondary structure formations to presently existing proteins (figure 12).
(ii) [GADV]-proteins contain not only amino acids necessary to form secondary structures ([A]: α-helix, [V]: β-sheet, [G]: β-turn), but also an amino acid, [D], with a carboxyl group to act as a catalyst (table 2).
(iii) Guanine is usually contained at the highest level at the first codon position of extant genes (figure 20A, B1). This suggests that the [GADV]-amino acids encoded by GNC are the most fundamental and important amino acids out of 20 kinds of natural amino acids. This feature might be maintained by post-code selection to keep correct reading frames.

Therefore, it can be deduced that even [GADV]-proteins, which have simple amino acid compositions, could catalyze the formation of peptide bonds between [GADV]-amino acids at a high probability. If the deduction were correct, [GADV]-proteins could synthesize similar [GADV]-proteins at a high probability even in the absence of genes. That is because only 4 kinds of amino acids are used in the [GADV]-proteins. In addition, properties of [GADV]-proteins synthesized should be similar to each other, since [V] with a hydrophobic side chain and [D] with a carboxylic group should locate at an inner and a surface parts of globular proteins in water respectively. Of course, the diversity of [GADV]-proteins with 100 residues is as high as about 10^{60}. But it is supposed that every 4 amino acid sequences would be detected at a probability of one time in proteins with only $4^4 = 256$ residues. Therefore, it can be easily imagined that middle-sized [GADV]-proteins composed of 256 amino acids should be similar to each other. Furthermore, the earliest [GADV]-proteins might be produced by aggregation of oligopeptides with 10–20 residues to eliminate the astronomic combination and to make the probability pretty high. This means that [GADV]-proteins could be pseudo-replicated to produce similar [GADV]-proteins in the absence of genes. By taking these points into consideration, we arrived at the [GADV]-protein world hypothesis on the origin of genes (figure 21) (Ikehara 2000). It has been widely believed that proteins can not be used as the first materials for creation of life, because proteins cannot be self-replicated, in spite that amino acids in proteins should be more easily synthesized than nucleotides in RNA and DNA on the pre-biotic earth. However, our novel hypothesis based on the characteristics of [GADV]-proteins would overcome the weak point of the protein world on the origin of life by the pseudo-replication of [GADV]-proteins in the absence of genes. Contrary to that, we found that there are several big defects in "RNA world hypothesis", although the hypothesis has been supported by many persons at present. The weak points of the RNA world hypothesis is described in the following section.

5.2 *Hypothesis on the origin of life provided by other researchers*

5.2a *RNA world hypothesis*: It is generally considered that replication of genetic materials is the most fundamental and important to life, wherein genetic information contained in the DNA is propagated and maintained by proteineous replication enzymes. As is well known, DNA usually does not have catalytic functions, whereas proteins cannot be used as genetic materials. Thus, DNA carrying genetic information cannot be replicated without proteins

and proteins cannot be reproduced without genes. Many persons have thought that the difficult problem on the origin of life might be solved, when catalytic activities were discovered in RNA having nucleotide sequences similar to DNA (Kruger *et al* 1982; Guerrier-Takada *et al* 1983). Based on the facts that RNA has not only genetic functions but also catalytic functions, the RNA world hypothesis has been provided, suggesting that RNA had been amplified by self-replication and increased their diversity in the RNA world on primitive earth (figure 22) (Gilbert 1986; Gesteland *et al* 1999).

But there are many defects in the RNA world hypothesis, as described below. (i) Nucleotides in RNA are organic compounds far more complex than amino acids in [GADV]-proteins. Thus, it would be apparently difficult to synthesize four nucleotides, A, U, G and C, than four [GADV]-amino acids under pre-biotic conditions. It is apparent from comparison the numbers of atoms in and of isomers of the nucleotides with those of the amino acids. (ii) Nevertheless, assume that the nucleotides could be synthesized under the pre-biotic conditions. But judging from the number of hydroxyl groups in the nucleotides, it would still be more difficult to synthesize RNA by joining them in the absence of effective catalysts than the peptide formation between amino acids. (iii) Furthermore, assume that RNA could be synthesized under the pre-biotic conditions. It would be difficult to self-replicate RNA, because RNA without any stable tertiary structure is required to exhibit a genetic function on nucleotide sequence as a template. Simultaneously, RNA must hold a stable tertiary structure to exhibit catalytic function on RNA. Therefore, it would be impossible to self-replicate

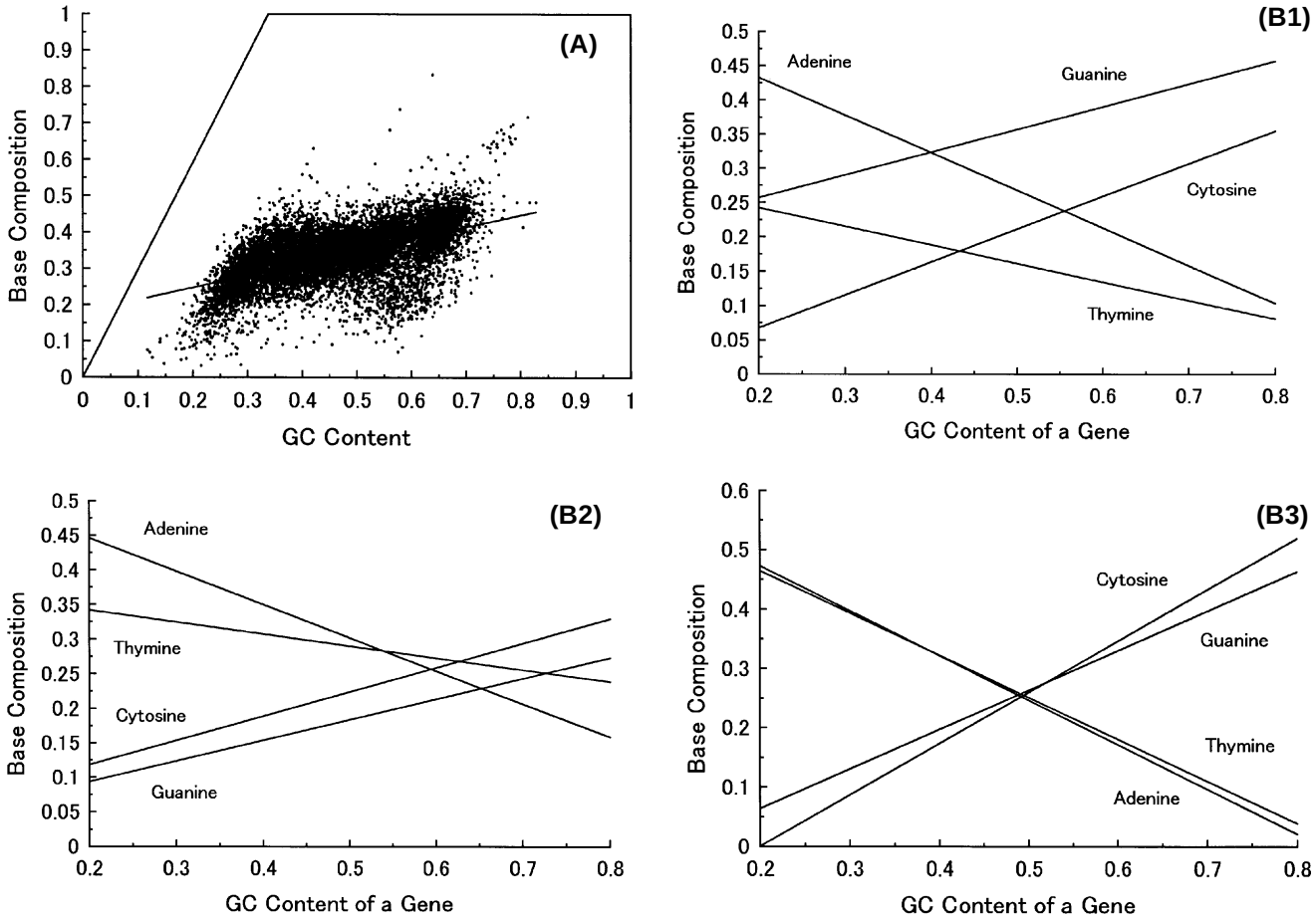

Figure 20. (**A**) Dependence of guanine composition at the first codon position (G1) upon GC content of a gene. The rectangular surrounding dots of G1 shows borders of the possible G1 composition, when all sequences in the sequence space could be used as functional genes. (**B**) Correlation of base compositions in the codon with GC content of a gene. The straight lines indicate average guanine, adenine, cytosine and thymine compositions at the first (**B1**), second (**B2**) and third (**B3**) base positions in the codon, respectively.

RNA in a usual sense due to the above self-contradiction. In fact, experimental results indicating that RNA molecules were actually self-replicated, have not been reported until now, in spite of many research works on RNA self-replication (Gesteland *et al* 1999). (iv) In addition to the above difficulties, there is another difficulty in the RNA world hypothesis. There should be no relationship between the ability for self-replication of RNA and the genetic function on RNA sequence for synthesis of a protein. Therefore, even though RNA was actually self-replicated on primitive earth, it is difficult to figure out that self-replicated RNA could simultaneously acquire any genetic information for protein synthesis. Thus, we consider that the RNA world was never realized on primitive earth (Ikehara 1999, 2000).

Although many researchers have discussed the origins of genes, the genetic code, proteins and life, they have treated it as four independent fundamental problems. Moreover, we have now recognized that there are big defects in their hypotheses on the origins of fundamental life systems, as described above. Contrary to that, it could be possible to explain comprehensively the four origins from a standpoint of the GNC-SNS primitive genetic code hypothesis (figure 23). Thus, I firmly believe that our hypotheses on the origins of genes and genetic code, proteins and life are far more reasonable than the hypotheses provided by other researchers.

Are our scenario with four stages really correct? To confirm this, we investigated further, whether it is possible to explain some properties of presently existing genes and extant proteins according to our hypotheses or not. I will explain about the results in the following section.

6. Evidences for our hypotheses on the fundamental life systems

6.1 *On the field for gene creation and the evolutionary directions of original genes and proteins*

If our hypotheses on the origin of genes (the GC-NSF(a) hypothesis and the $(SNS)_n$ hypothesis) were really correct, genes should be created originally as GC-rich genes, and genes homologous with the ancestors would be produced during an evolutionary process, as GC contents of the genes gradually decreased under AT mutation pressure (figure 24). Thus, according to our hypotheses, proteins could be originally created by expression of the ancestor genes with high GC contents, and the original proteins should evolved to homologous proteins encoded by the genes with low GC contents (or AT-rich genes). In addition, characteristics of the ancestor proteins should be maintained in conserved regions among homologous proteins. Therefore, evolutionary directions of genes and proteins could be deduced by investigating the properties, or conserved amino acids and non-conserved amino acids, which are observed among homologous proteins after alignments of the proteins (figure 25). To confirm this actually, the amino acid sequence of *P. aeruginosa* gyrase A (GyrA), which is encoded by a GC-rich gene, was compared with other homologous GyrA proteins, which are encoded by genes with lower GC contents than the *P. aeruginosa gyrA* gene. From the results, it was found that the contents of SNS-encoding amino acids (SNS-AAs) in conserved regions were similarly as high as to SNS-AA contents in non-conserved regions of *P. aeruginosa* GyrA protein (figure 26A). It was also confirmed

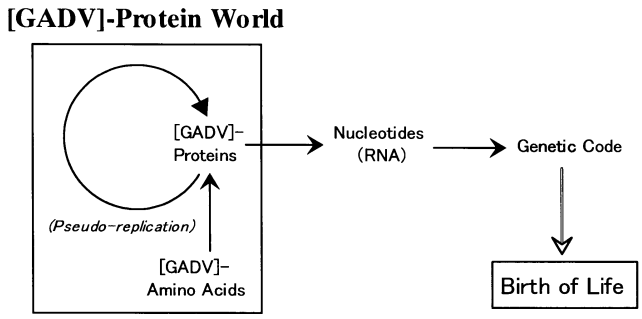

Figure 21. [GADV]-protein world hypothesis on the origin of life. The hypothesis anticipates that life originated from [GADV]-protein world, where [GADV]-proteins were amplified by themselves through pseudo-replication. It is supposed that the simple amino acid composition composed of only four amino acids (Gly [G], Ala [A], Asp [D] and Val [V]) made it possible to pseudo-replicate [GADV]-proteins in the absence of any genetic function.

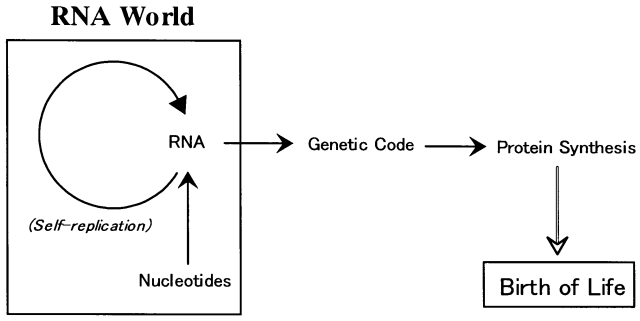

Figure 22. RNA world hypothesis on the origin of life. The RNA world hypothesis is based on self-replication of RNA, which has been proposed to explain the development of life on the earth. The unanticipated discovery of RNA catalysts or ribozymes initiated extensive discussion of the role of RNA on the origin of life. But, we think that there are many weak points in the RNA world hypothesis to be impossible to solve, as described in this review.

that SNS-AA contents decreased gradually in non-conserved regions of other GyrA proteins encoded by genes with lower GC contents, as GC content of the genes decreased (figure 26A).

The SNS-AA contents in conserved and in non-conserved regions were also investigated by using SpoT/RelA proteins (figure 26B), glutamine synthetase (GlnA), *a*-subunit of RNA polymerase (RpoA) and other 10 kinds of homologous proteins. Similar results to the case of GyrA were obtained in all cases examined (data not shown). The results described above clearly indicate that water-soluble globular proteins, such as GyrA and SpoT/RelA, were generally created as ancestor proteins encoded by GC-rich genes and evolved unidirectionally to the proteins encoded by genes with lower GC contents, as was expected.

6.2 *Process of protein formation*

Assuming that proteins were originally produced in the field expected by our hypothesis on the origin of proteins, new proteins should be created by random joining of amino acids restricted in SNS 0th-order structure and in GC-NSF(a) 0th-order structure. The SNS and GC-NSF(a) 0th-order structures are the amino acid compositions determined by SNS-repeating sequences ($(SNS)_n$) and nonstop frames on antisense strands of GC-rich genes (GC-NSF(a)) respectively (figure 19). The GNC 0th-order structure is omitted only to simplify the discussion in this section. If new proteins were actually produced by random joining of the amino acids, frequencies of two neighbouring amino acids, which are observed in presently existing proteins, should be coincident with those obtained by multiplication of the two amino acid compositions. The results obtained using *H. influenzae* genome data clearly indicate that 400 spots, representing all combinations of two neighbouring amino acids, were closely distributed around the linear line with slope 1 (figure 27). Genome data of *H. pylori*, *E. coli*, *M. genitarium* and *B. subtilis* gave similar results to the case of *H. influenzae*. These facts indicate that proteins would be fundamentally created by random joining of amino acids in the amino acid compositions, which are restricted by the SNS and GC-NSF(a) 0th-order structures.

6.3 *Simulation of origins and evolutions of genes and proteins with a computer*

As was explained above, it seems to us that our hypotheses on the origins of genes, the genetic code, proteins and life are really correct. If that is true, it is considered that genes were produced as GC-rich genes and decreased their GC contents when mutations accumulated on the nucleotide sequences. Then every protein encoded by genes, which were produced by accumulation of mutations on the sequences, should satisfy the six conditions

Our hypothesis	Origin of	Other researcher's hypothesis
[GADV]-Protein World	Life	RNA World
GC-NSF(a) (SNS)n	Gene	Gene Duplication Exon Shuffling
GNC Primitive Code SNS Primeval Code	Genetic Code	RNY Code WWW Code Mitochondrial-type Code
GC-NSF(a) 0^{th}-order Structure SNS 0^{th}-order Structure	Protein	Sequence theory Structure theory

Figure 23. Comparison of our hypotheses on the origins of genes, genetic code, proteins and life, which are involved in fundamental systems of life, with those postulated by other researchers. Our hypotheses are interrelated each other, mainly based on the GNC-SNS genetic code hypothesis. Contrary to that, those postulated by other researchers have been rather independently discussed without any correlation among four fundamental systems of life.

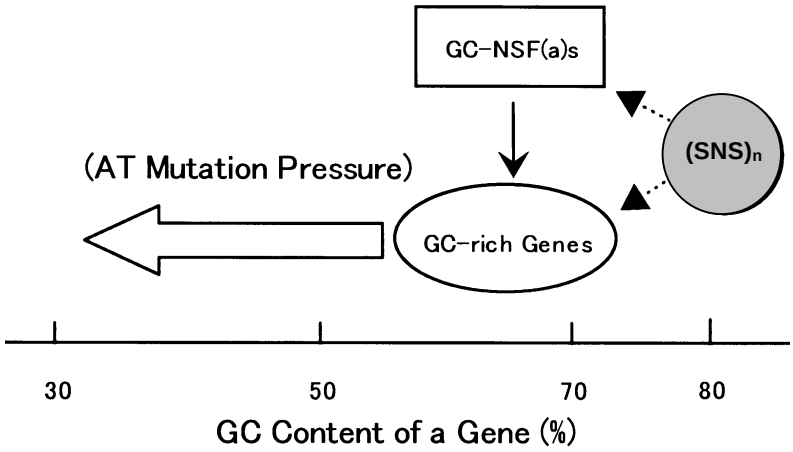

Figure 24. Model of gene evolution presented in this review. According to the model, genes should originated from SNS-repeating sequences, $(SNS)_n$, in the ancient days and GC-NSF(a) in the present days, as GC-rich genes. Thus, generally, genes must be unidirectionally propagated from GC-rich ancestral genes to AT-rich genes under AT-mutation pressure.

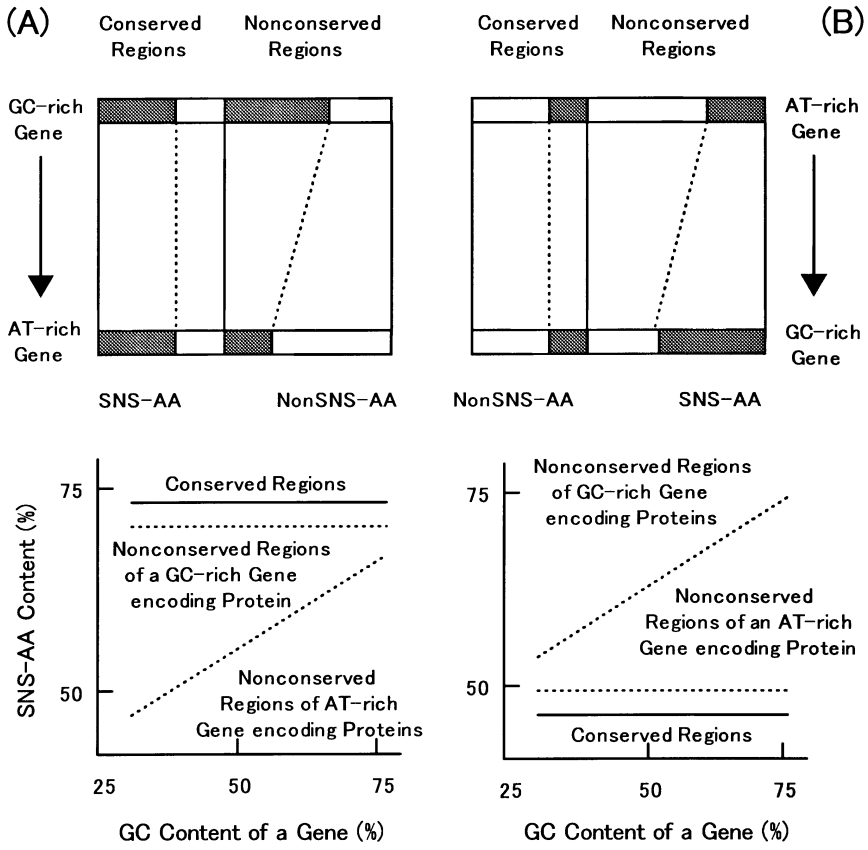

Figure 25. Deduced change of SNS-encoding amino acid (SNS-AA) contents in conserved and non-conserved regions of homologous proteins occurring in a process of gene evolution. (**A**) Change of SNS-AA contents (shaded boxes) is deduced according to our evolution models of genes (figure 24) and proteins (figure 19), in which genes have been evolved unidirectionally from GC-rich to AT-rich genes. In this case, SNS-AA contents in regions conserved among homologous proteins keep them high during the gene evolution, while SNS-AA contents in nonconserved regions of proteins encoded by AT-rich genes gradually decrease as the GC content of a gene decreases. (**B**) Change of SNS-AA contents is deduced according to a model, in which genes have been evolved unidirectionally but from AT-rich to GC-rich genes. In this case, SNS-AA contents (shaded boxes) in regions conserved among homologous proteins remains low during the gene evolution, while SNS-AA contents in nonconserved regions of proteins encoded by GC-rich genes gradually increase as the GC content of a gene increases.

(hydrophobicity/hydrophilicity, *a*-helix, *b*-sheet and *b*-turn formabilities, acidic amino acid and basic amino acid contents), which are required to form appropriate tertiary structures of proteins. If the speculation is correct, changes of base compositions in the codon positions of genes, as well as of amino acid compositions of proteins, could be simulated according to our hypotheses, expecting the field of gene production and the evolutionary processes of genes and proteins.

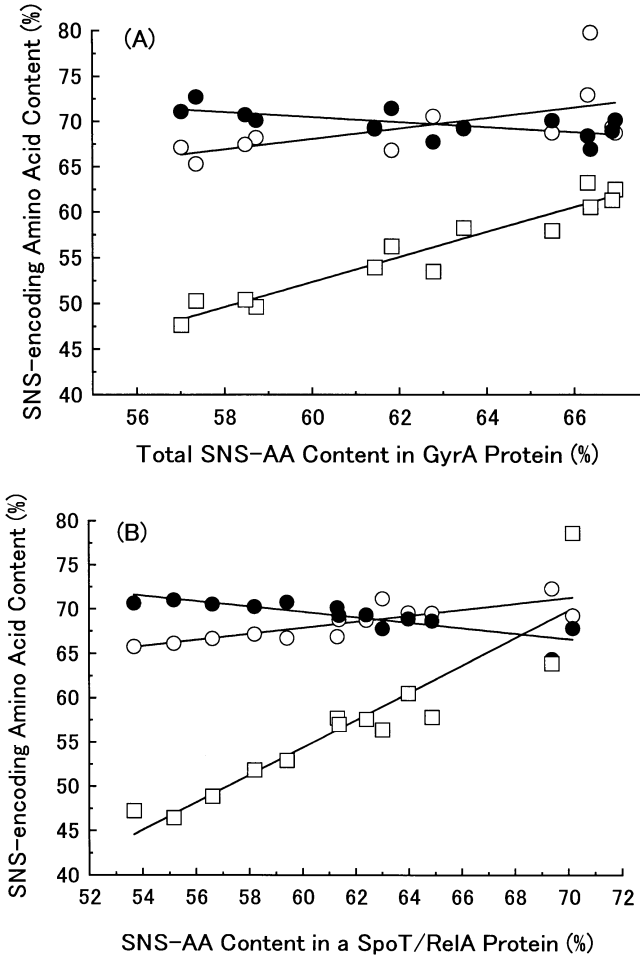

Figure 26. SNS-AA contents in conserved and nonconserved regions among homologous proteins. SNS-AA contents in both conserved (open circles) and non-conserved (closed circles and open squares) regions were obtained after alignments of GyrA protein encoded by GC-rich *Pseudomonas aeruginosa gyrA* gene (**A**); and of RelA/SpoT protein encoded by GC-rich *Streptomyces coelicolor* gene (**B**); with the corresponding homologous proteins from 13 other microbial taxa. These results clearly indicate that both GyrA genes and RelA/SpoT genes changed unidirectionally from GC-rich ancestral genes to AT-rich genes under AT-mutation pressure, as was expected from our hypothesis on gene evolution (figure 25).

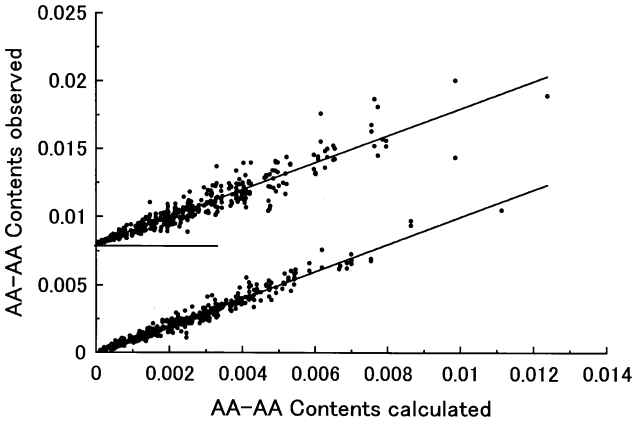

Figure 27. Comparison of frequencies of two neighbouring amino acid residues observed on protein sequences from bacterial genomes with those estimated by multiplication of two amino acid compositions of proteins encoded by bacterial genomes of *H. influenzae* (lower plots) and *H. pylori* (upper plots). Points of 400 combinations of two neighbouring amino acids should distribute around the linear line with a slope of 1, if proteins were produced through random peptide formation of amino acids restricted in protein 0th-order structures, such as SNS- and GC-NSF(a) 0th-order structures. Upper plots *H. pylori* proteins were shifted by addition of 0·008 to the raw data to make two sets of data to compare easily.

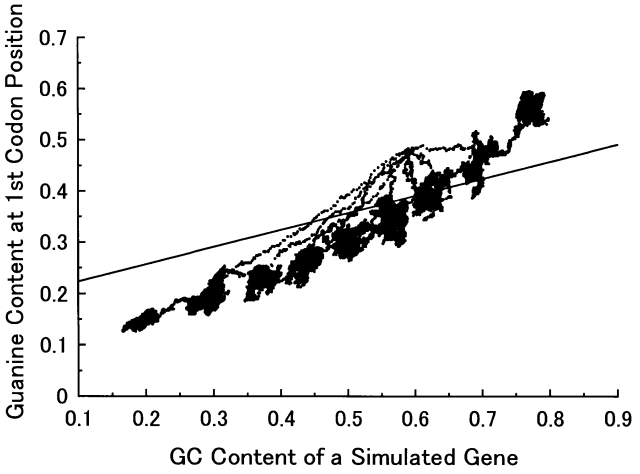

Figure 28. Computer simulation of the evolution of genes derived from a GC-rich ancestral gene. The simulation was carried out by using a GC-NSF(a) of a *M. tuberculosis* GC-rich gene composed of 1,500 bases as an ancestor gene. The simulation of gene evolution was carried out under 10 different mutation pressures. Only guanine composition at the first codon position (G1) out of twelve base compositions was largely deviated from the approximated linear line drawn using seven microbial genome data. Therefore, G1 composition is only given as an example of the results of the gene simulation in this figure.

We actually carried out the simulation of genes by using a GC-NSF(a) or an antisense sequence of a *M. tuberculosis* GC-rich gene, which is composed of about 1,500 bases (about 500 codons). Before the simulation, we had confirmed that the protein encoded by the hypothetical ancestor gene satisfy the six conditions required to form water-soluble globular proteins. Next, mutations were introduced onto the nucleotide sequence at a probability of 1% at every base position. Thus, the mutation probability corresponds to 15 base substitutions on the gene or to about 5 amino acid replacements in the protein. Then, if a simulated protein did not satisfy one or more of the six conditions, or if stop codon(s) appeared in the frame of translation, it was removed from the simulation as an inactive gene or a mutated protein. And the procedure for introduction of mutations was repeated from one step before. These steps were repeated until getting 1,000 hypothetical proteins satisfying the six conditions for globular protein formation under a mutation pressure. The resulting simulations, which were carried out under 10 different mutation pressures, are given in figure 28. A linear line, which was obtained by the least square approximation of G compositions at the first codon position (G1) of 7 extant microbial genomes, is drawn in figure 28 to compare it with the simulated results. Twelve dependencies of base compositions on GC content of a gene were obtained from the simulation, because 4 kinds of bases (A, T, G and C) exist at every three codon positions. Although the simulated base changes reproduced well those of actual genes in 7 microbial genomes upon the change of the GC content, only G1 was considerably deviated from the linear line (figure 28). To know the causes of the deviation, simulation was similarly carried out again as 200 amino acid residues from N-terminal end of the ancestor protein composed of about 500 amino acids were remained unchanged. The size of the conserved region was taken in the simulation, after considering the fact that about 40% amino acids are generally conserved among homologous proteins. From the results, it was found that the simulated G1 also expectedly reproduced the change of G1 in presently existing genes (figure 29). In addition, amino acid compositions of the simulated proteins were also fairly coincident with those of extant proteins encoded by microbial genomes except of arginine, lysine and some other amino acids (figure 30).

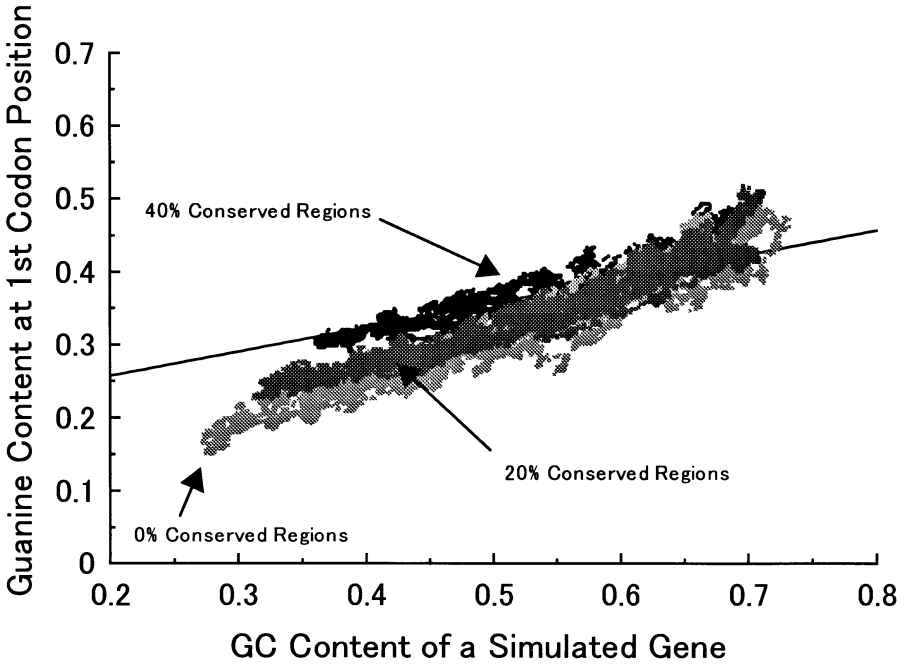

Figure 29. As seen in the results of figure 28, where the simulation was carried out under the condition without any conserved region among evolved proteins, guanine composition at the first codon position was largely deviated from the line drawn using seven microbial genome data. In this figure, the simulation was carried out using a *M. tuberculosis* gene (sense sequence) as an ancestor gene under the conditions with 0, 20 and 40% conserved regions. From the results, it was confirmed that G1 composition became closer and closer to the line, as the ratio of conserved regions increased from 0 to 40%.

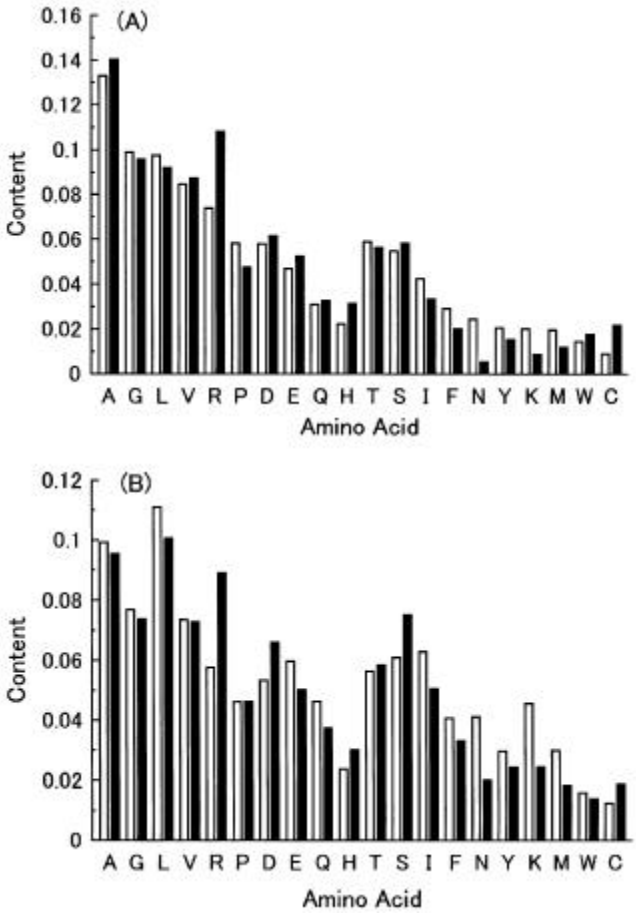

Figure 30. Comparison of average amino acid compositions of proteins obtained (open bars) from *M. tuberculosis* genome data (average GC content = 65·6%) (**A**), and *E. coli* genome data (average GC content = 50·8%) (**B**), with those of hypothetical proteins (closed bars) obtained, when conserved regions were maintained at 40% during the gene simulation, as given in figure 29. The average GC contents of the simulated genes were 65·6% (**A**); and 50·5% (**B**), respectively.

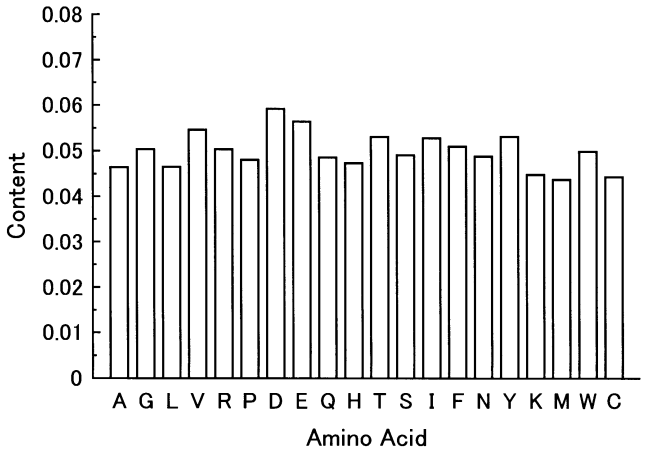

Figure 31. Amino acid compositions of hypothetical proteins, which were generated as proteins with amino acid compositions specified by 20 independent random numbers and simply selected only using the six conditions for formation of globular proteins. This shows that all 20 amino acids in proteins have similar abilities for tertiary structure formation of globular proteins. Therefore, it is evident that specific distributions of amino acid compositions in *M. tuberculosis* and *E. coli* proteins as seen in figure 30, have been determined during evolutionary processes of proteins from ancestor proteins encoded by GC-rich genes.

To confirm that the distributions of amino acid compositions have not been determined simply by characteristics of 20 natural amino acids, amino acid compositions were simply generated with 20 random numbers by a computer, independently on an ancestor gene and its evolutionary pathway. Next, when proteins with an amino acid composition specified by the 20 random numbers satisfied the six conditions for globular protein formation, the proteins were selected out as functional proteins. Average amino acid compositions of the selected hypothetical proteins were drawn as a bar chart in figure 31. The result shows that all amino acids were contained at similar ratios in the selected proteins. It is in marked contrast to the fact that the amino acids, such as Leu, Ala, Gly, Val, Ser and Ile, are usually detected in extant proteins at high frequencies, but the amino acids, such as His, Met, Trp and Cys, were usually observed at amounts less than other amino acids. Therefore, the results described above indicate that the average amino acid compositions of presently existing proteins have been determined by the facts that genes had been originally created as GC-rich genes and have evolved to the direction lowering the GC content, not by the characteristics of the amino acids (figures 29 and 30). Simultaneously, it means that proteins have been produced by translation of genetic information of genes appeared during the evolutionary process. These results also support our hypotheses on the origins of genes, genetic code and proteins and their evolutionary processes.

7. Conclusion

As discussed in the above sections, we have provided four novel hypotheses on the origins of genes, the genetic code, proteins and life, by using chiefly the six conditions for the formation of globular proteins. As a matter of course, we are confident in the four origins on the fundamental life systems, which can be reasonably and comprehensively explained by our hypotheses. Contrary to that, it is difficult to understand the origins by using the hypotheses provided by other researchers. The main reasons, why the four problems could not be well inter-

preted until now, in spite of the many efforts by other researchers, would be attributed to the facts that they have discussed the problems rather independently. Therefore, although the main current of the origin of life would be "RNA-world hypothesis" at present, we also believe firmly that our [GADV]-protein world hypothesis is far more reasonable to explain the origin of life than the RNA-world hypothesis. In addition, we hope that it makes possible to interpret various properties of extant genes and proteins, and even evolutionary pathways of the present metabolism, by reconsideration of the fundamental life systems on genes, the genetic code and proteins from the standpoint of our hypotheses. In fact, studies are in progress in our laboratory to solve these problems on fundamental systems of the present life.

Acknowledgements

I thank Dr Vidyanand Nanjundiah and Dr Apoorva Patel for recommending the submission of this manuscript to *J. Biosci*., as an English version of the paper (*Viva Origino* 2001 **29** 66–85) written in Japanese. I also thank Dr Noriko Fujii (Research Reactor Institute Kyoto University, Editor-in-Chief of *Viva Origino*) for approval of submitting the original paper to *J. Biosci*., as a modified English version.

References

Dill K A 1990 Dominant forces in protein folding; *Biochemistry* **29** 7133–7155

Gesteland R F, Cech T R and Atkins J F 1999 *The RNA world* (New York: Cold Spring Harbor Laboratory Press)

Gilbert W 1986 The RNA world; *Nature* (*London*) **319** 618

Gilbert W, de Souza S J and Long M 1997 Origins of genes; *Proc. Natl. Acad. Sci. USA* **94** 7698–7703

Guerrier-Takada C, Gardiner K, Marsh T, Pace N and Altman S 1983 The RNA moiety of ribonuclease P is catalytic subunit of the enzyme; *Cell* **35** 849–857

Ikehara K 1998a Origin and evolution of the genetic code (based on the novel SNS hypothesis) (in Japanese); *Seibutukagaku* **50** 44–54

Ikehara K 1998b A possible evolutionary pathway of the genetic code deduced from the SNS hypothesis; *Viva Origino* **26** 311–320

Ikehara K 1999 Is the RNA-world hypothesis of prebiotic evolution correct? (life was originated from the prebiotic protein world!) (in Japanese); *Seibutukagaku* **51** 43–53

Ikehara K 2000 Life was born from proteins!? – [GADV]-protein world hypothesis – (in Japanese); *Kagaku* **55** 14–19

Ikehara K, Amada F, Yoshida S, Mikata Y and Tanaka A 1996 A possible origin of newly-born bacterial genes: significance of GC-rich nonstop frame on antisense strand; *Nucleic Acids Res*. **24** 4249–4255

Ikehara K and Okazawa E 1993 Unusually biased nucleotide sequences on sense strands of *Flavobacterium sp*. genes produce nonstop frames on the corresponding antisense strands; *Nucleic Acids Res*. **21** 2193–2199

Ikehara K, Omori Y, Arai R and Hirose A 2002 A novel theory on the origin of the genetic code: a GNC-SNS hypothesis; *J. Mol. Evol*. (in press)

Ikehara K and Yoshida S 1998 SNS hypothesis on the origin of the genetic code; *Viva Origino* **26** 301–310

Jimenez-Sanches A 1995 In the origin and evolution of the genetic code; *J. Mol. Evol*. **41** 712–716

Kruger K, Grabowski P J, Zaug A J, Sands J, Gottschling D E and Cech T R 1982 Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*; *Cell* **31** 147–157

Miller S L and Orgel L E 1975 *The origin of life on the earth* (New York: Prentice-Hall)

Ohno S 1970 *Evolution by gene duplication* (Heiderberg: Springer)

Osawa S 1995 *Evolution of the genetic code* (Oxford: Oxford University Press)

Stryer L 1988 *Biochemistry* 3rd edition (New York: W H Freeman)

Sueoka N 1988 Directional mutation pressure and neutral molecular evolution; *Proc. Natl. Acad. Sci. USA* **85** 2653–2657

Voet D, Voet J G and Pratt C W 1999 *Fundamentals of biochemistry* (New York: John Wiley)

MS received 29 October 2001; accepted 6 February 2002

Corresponding editor: VIDYANAND NANJUNDIAH