THE CHEMICAL RECORD

Possible Steps to the Emergence of Life: The [GADV]-Protein World Hypothesis

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

Received 18 January 2005; Revised 24 February 2005; Accepted 22 February 2005

ABSTRACT: Based on the fact that RNA has not only a genetic function but also a catalytic function, the RNA world theory on the origin of life was first proposed about 20 years ago. The theory assumes that RNA was amplified by self-replication to increase RNA diversity on the primitive earth. Since then, the theory has been widely accepted as the most likely explanation for the emergence of life. In contrast, we reached another hypothesis, the [GADV]-protein world hypothesis, which is based on pseudo-replication of [GADV]-proteins. We reached this hypothesis during studies on the origins of genes and the genetic code, where [G], [A], [D], and [V] refer to Gly, Ala, Asp, and Val, respectively. In this review, possible steps to the emergence of life are discussed from the standpoint of the [GADV]-protein world hypothesis, comparing it in parallel with the RNA world theory. It is also shown that [GADV]-peptides, which were produced by repeated dry-heating cycles and by solid phase peptide synthesis, have catalytic activities, hydrolyzing peptide bonds in a natural protein, bovine serum albumin. These experimental results support the [GADV]-protein world hypothesis for the origin of life. © 2005 The Japan Chemical Journal Forum and Wiley Periodicals, Inc. Chem Rec 5: 107–118; 2005: Published online in Wiley InterScience (www.interscience.wiley.com) DOI 10.1002/tcr.20037

Key words: origin of life; [GADV]-protein world hypothesis; RNA world hypothesis; $(SNS)_n$ primitive genes; $(GNC)_n$ primeval genes; GNC–SNS primitive genetic code hypothesis

Introduction

Since the first complete microbial genomic sequence of *Haemophilus influenzae* was published from the Institute for Genomic Research (TIGR) in 1995, more than 230 genomic sequences of organisms, such as bacteria, archaea, and eukaryotes including human, have been already determined. With the accumulation of the genomic sequences in public databases, studies on the last common ancestor of all extant biological species have been carried out through comparative analyses of the completely sequenced cellular genomes and their proteomes, and several useful results have been obtained so far.[1–3] In contrast to that, we have investigated a scenario from abiotic syntheses of simple organic compounds to the first form of life,[4,5] which corresponds to events in a past era much older than that of the appearance of the last common ancestor.

The question of what caused changes from inorganic compounds to living matter on the primitive earth has been a scientific concern for about two thousand years. Many hypotheses for the origin of life have been proposed based on

▸ ***Correspondence to***: Kenji Ikehara; e-mail: ikehara@cc.nara-wu.ac.jp

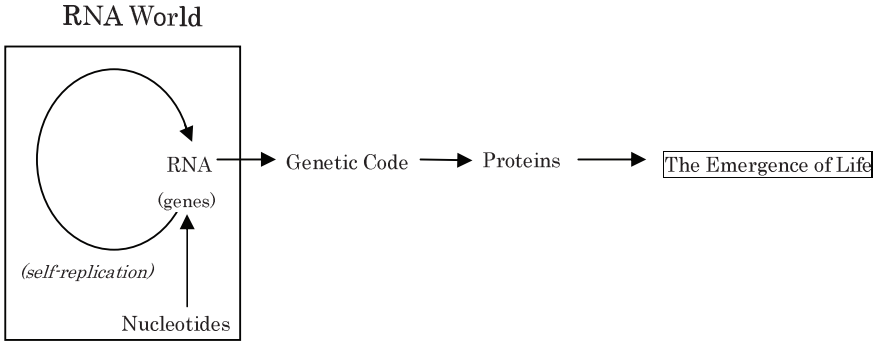

Fig. 1. RNA world hypothesis for the origin of life. The theory assumes that RNA with both genetic information and catalytic activity would self-replicate and produce diverse RNA molecules, and that accumulated RNA would create genetic code and proteins, leading to the emergence of life.

quite different viewpoints. Even the places from where it is proposed life first emerged on earth are quite different from each other.[6] For example, two concepts are that (i) life was created in a prebiotic soup on earth[7] or that (ii) life emerged from hydrothermal vents in the ocean.[8] Even a proposal that (iii) the first life forms arrived on earth from space has been presented, based on the detection of organic compounds, such as amino acids, in meteorites.[9]

It is well known that DNA generally does not exhibit any catalytic functions, whereas proteins cannot be used as genetic materials. Thus, DNA carrying genetic information cannot be replicated without proteins, whereas proteins cannot be reproduced without genes. This presents a so-called "chicken and egg relationship" in the present life systems. However, about twenty years ago, it was found that RNA has some catalytic activities.[10,11] Based on the fact that RNA has not only genetic functions but also catalytic functions, Gilbert proposed the RNA world theory on the origin of life, suggesting that RNA had been amplified by self-replication and increased RNA diversity in the world, and that life originated from the RNA world (Fig. 1).[12,13] The idea was accepted by many investigators as a trump card for solving the "chicken and egg problem" of the origin of life. Thus, the RNA world hypothesis has become widely accepted among the theories on the origin of life.

In contrast, about 10 years ago, we began research on the origin of genes based on the following considerations. (i) If new genes could be produced on earth at present, from what kind of field on the DNA sequences could new genes have been created under the universal genetic code? (ii) From what kind of DNA sequences were genes produced on the primitive earth?

From analyses of microbial genes and proteins obtained from the GenomeNet Database, we found that new genes could be produced from non-stop frames on antisense sequences of microbial GC-rich genes (GC-NSF(a)).[14] As a

Kenji Ikehara is a professor in the Department of Chemistry of Nara Women's University. He was born in 1944, graduated from the Department of Industrial Chemistry, Kyoto University in 1968 and received his B.Eng. and D.Eng. degrees from Kyoto University. He has been engaged in studies on the fundamental properties of microbial genes, the genetic code, proteins, and metabolism. As a result of these studies, he proposed the [GADV]-protein world hypothesis for the origin of life. ■

next step, we performed an analysis on the origin of the genetic codes. Consequently, we have reached a GNC–SNS primitive genetic code hypothesis, suggesting that the universal genetic code originated from the GNC code through to the SNS code, where N and S represent either of four bases (G, C, A, and T or U) and G or C, respectively.[15–18] Based on the origins of genes and the genetic code, we have also proposed a novel hypothesis, the [GADV]-protein world hypothesis, which could reasonably explain the emergence of life.[4,5] The [GADV]-protein world is a world composed of four amino acids (Gly: [G], Ala: [A], Asp: [D], and Val: [V]). In this review, we will present an idea concerning possible steps to the emergence of life, based on the [GADV]-protein world hypothesis.

GC-NSF(a) Hypothesis for New Gene Creation Under the Universal Genetic Code

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 were largely changed as the GC contents of the genes increased or decreased under GC- or AT-mutation pressure.[5,18] However, 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), should be invariable against the differences of GC contents among genes followed by those of amino acid compositions among proteins, since proteins must generally be folded into appropriate three-dimensional structures in every microbial cell carrying a chromosomal DNA even 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 by the indexes of 20 amino acids from Stryer's textbook for the analyses of hydropathy and of secondary structure formations.[19] From analyses of the six indexes of microbial proteins, it was discovered that all indexes are, as expected, almost constant against the differences of the GC contents of genes obtained from seven genomes of bacteria and archaea.[18] This means that when we judge whether polypeptide chains can be folded into water-soluble globular structures required to exhibit enzymatic functions, the six structural indexes of water-soluble globular proteins should be treated as necessary conditions. Thus, we first searched for a field by using the conditions where new genes could be created under the universal genetic code. For this purpose, we investigated the six structural indexes of hypothetical proteins encoded by possible 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 cells at a high probability.[14] Moreover, the probability (pNSF) at which no stop codon appears in a reading frame increases abruptly beyond about 60% GC content, which is caused by unusually biased base compositions at three codon positions.[5,14,20] Hereafter, we designate the nonstop frame (NSF) on the antisense strand of the GC-rich gene as GC-NSF(a). Proteins encoded by hypothetical GC-NSF(a) genes also 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 do those of actual proteins.[14] Thus, we assert that the GC-NSF(a)s easily found on GC-rich microbial genomes must be the field where new genes could be produced on earth at present (Fig. 2).

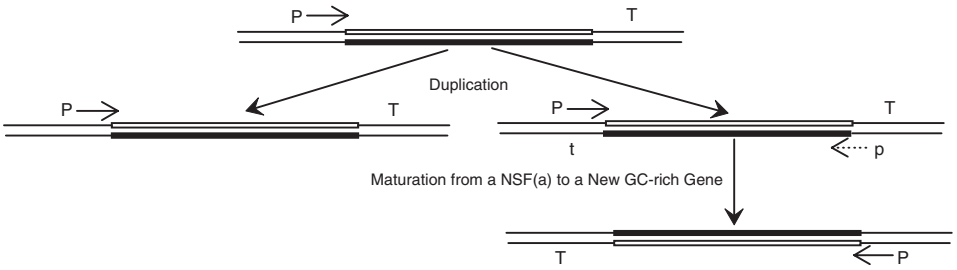

Fig. 2. GC-NSF(a) hypothesis for the origin of genes, predicting that new genes originate from non-stop frames on antisense strands of GC-rich genes (GC-NSF(a)s). It is supposed that a GC-NSF(a) gene would evolve to get greater activity by accumulating necessary mutations if a protein were expressed from a latent promoter (p) and the GC-NSF(a) possessed even the faintest activity necessary to live. The new GC-rich gene could then be propagated to other bacteria having more AT-rich genes, and the GC content of the gene would decrease under unidirectional AT-mutation pressure (gene evolution). Uppercase letters, "P" and "T," and lowercase letters, "p" and "t," refer to functional promoters and effective transcription terminators, and latent promoters and inefficient terminators, respectively. Long rectangles and bold lines indicate functional genes and cryptic genes, respectively.

Origins of Genes and the Genetic Code on the Primitive Earth

$(SNS)_n$ Hypothesis for the Origin of Genes and SNS Primitive Genetic Code Hypothesis

As described above, it is expected that the GC-NSF(a)s would be used as a field for the creation of new genes under the universal genetic code.[14,20] It was also found 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.[5,18] This implies that even randomly repeating SNS sequences could be utilized at a high probability as functional genes encoding water-soluble globular proteins.

To confirm this, SNS compositions in the codon were created by using computer-generated random numbers. 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 amino acids as described above. The computer-generated SNS code satisfied the six conditions when the contents of G and C at the first codon position were around 55% and 45%, respectively, and when every four base was contained at a ratio of about one-quarter at the second codon position.[5,15,16,18] 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 is 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 structures.[5,15,16,18] 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). This indicates that SNS repeating sequences, $(SNS)_n$, could be used as primitive genes ($(SNS)_n$ primitive gene hypothesis).

The hypothesis means that proteins encoded by $(SNS)_n$ primitive genes should be produced under the SNS code because the SNS genetic code is sufficient to translate SNS repeating sequences. Thus, we have also provided an SNS primitive genetic code hypothesis.[5,15,16,18]

$(GNC)_n$ Hypothesis for the Origin of the Most Primitive Genes and GNC Primeval Genetic Code Hypothesis

The SNS code is composed of 10 amino acids encoded by 16 codons (Fig. 3). Thus, it must have been almost impossible to create an SNS code at one stroke on the primitive earth. To solve this problem, we investigated what kind of repeating

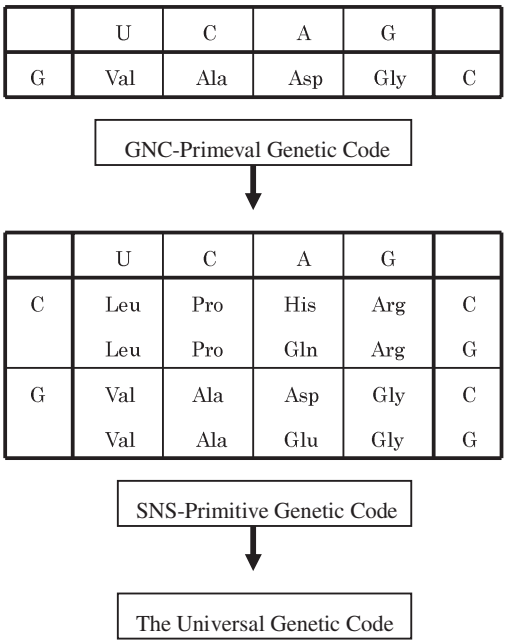

Fig. 3. GNC–SNS primitive genetic code hypothesis, which we have proposed. The hypothesis postulates that the universal genetic code originated from the GNC primeval code (4 codons and 4 amino acids) through to the SNS primitive code (16 codons and 10 amino acids). "N" and "S" mean either of four bases (A, U, G, and C) and G or C, respectively.

sequences that are much simpler than SNS repeating sequences could code for water-soluble globular proteins. For this purpose, four conditions (hydropathy, α-helix, β-sheet, and β-turn formations) out of the six conditions for globular protein formation were used to determine which base sequences could encode water-soluble globular proteins. The results indicated that four [GADV]-amino acids encoded by GNC repeating sequences ($(GNC)_n$) satisfied well the four structural conditions.[5,18] Here, it can be concluded that GNC random sequences, $(GNC)_n$, could be used as the most primeval genes on the primitive earth before the emergence of $(SNS)_n$ primitive genes.[5,18] The GNC random sequences are symmetric between sense and antisense sequences, meaning that both sequences could be used for similar but not the same information for the synthesis of [GADV]-proteins. This implies that the most primitive genes, $(GNC)_n$, are favorable for pseudo-replication of [GADV]-proteins under the most primitive genetic code.

Miller's experiments[21] indicating that four amino acids ([G, A, D, and V]) could easily be formed on the primitive earth support the above $(GNC)_n$ primeval gene hypothesis and the GNC primeval genetic code hypothesis, which we have proposed. Recently, Trifonov provided a temporal order of appearance of amino acids, as G, A, D, V, P, S, E, (L, T), R, (I, Q, N), H, K, C, F, Y, M, W, based on 60 different criteria

each offering temporal order.[22] The order is coincident to a high degree with the GNC–SNS primitive genetic code hypothesis, also supporting the hypothesis on the origin of the universal genetic code.

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 appear in the random sequences at extremely high frequencies. At present, however, it is unknown what makes it possible to enable group coding.

Origin of Proteins

The tertiary structure of a protein is determined by an amino acid sequence, which is encoded by a gene and the genetic code. This means that the origins of genes and the genetic code should be intimately related to the origin of proteins, and that proteins would coevolve with genes and the genetic code (Fig. 4). From this consideration, it is suggested that it would be possible to clarify the origins of proteins or mechanisms producing new proteins by applying our hypotheses for the origins of genes and the genetic code, as described above.

We next give consideration to the fields from which globular proteins were produced on the primitive earth.[5] In the sections above, we have used amino acid compositions instead of amino acid sequences for the structural analyses of hypothetical proteins. This strategy was effective for judging which proteins having an amino acid composition could be folded into appropriate tertiary globular structures in water. This means that proteins would generally be produced by the random assembly of amino acids encoded by GNC and SNS genetic codes. Thus, we would like to call these novel ideas on the origin of proteins the GNC and SNS 0th-order structure hypotheses because the GNC and SNS primitive genetic codes would specify amino acid compositions but not amino acid sequences (Fig. 5). In other words, the GNC and the SNS 0th-order structures were utilized in the early and later periods,

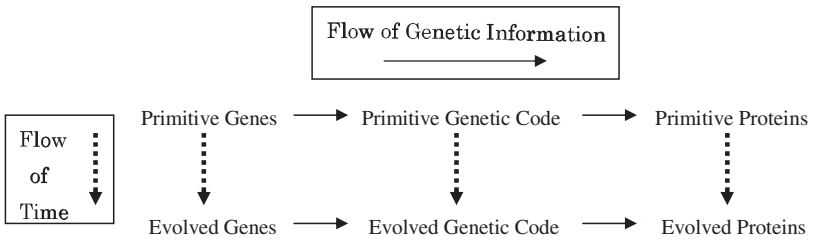

Fig. 4. A protein having an amino acid sequence is synthesized according to a nucleotide sequence or a gene, which is given as a series of triplet codons (three nucleotide sequences; genetic code). Formation of the tertiary structure of a protein is based on the amino acid sequence specified by the genetic code. Therefore, genes and the genetic code should be intimately related to proteins. These considerations suggest that genes, genetic code, and proteins coevolved and, therefore, the origin of proteins or the mechanisms producing proteins could be clarified according to our hypotheses for the origins of genes and the genetic code.

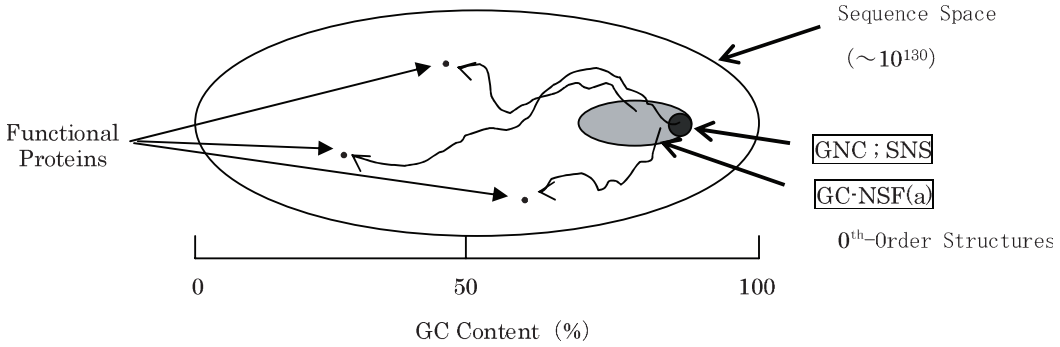

Fig. 5. GNC-, SNS-, and GC-NSF(a)-0th order structure hypotheses for the origin of proteins. According to the 0th order hypotheses, primitive proteins originated from random sequences (closed circles) composed of particular amino acids, which are specified by the GNC and SNS primitive genetic codes. At present, new proteins could be derived from proteins encoded by ancestral GC-NSF(a) genes (shaded ellipsoids), a modified form of SNS repeating sequences, $(SNS)_n$. The hypotheses suggest that new proteins could be produced by the random assembly of amino acids restricted in the 0th order structures.

respectively, to create new proteins efficiently on the primitive earth.[5] We would like to emphasize here that, if necessary, the amino acid compositions specified by the GC-NSF(a) hypothetical genes (GC-NSF(a) 0th-order structure or a modified form of the SNS 0th-order structure) must be used to produce new proteins on earth today (Fig. 5).[5]

As a matter of course, genes, the genetic code, and proteins are the most fundamental and important functions required for life. As described above, the GNC code should be the genetic code used in the most primitive life since the code would be the simplest one of all that has appeared on earth. This also means that the simplest set of amino acids, which were used to produce functional globular proteins on the primitive earth, would be [GADV]-amino acids encoded by the GNC code. Taking these facts into consideration, we have also provided a novel hypothesis for the origin of life, as described in the next section.[4,5]

[GADV]-Protein World Hypothesis for the Origin of Life

According to the GNC primeval genetic code hypothesis, the most primitive proteins should be composed of only four [GADV]-amino acids. Moreover, three of the four amino acids possess the capabilities required for the formation of respective secondary structures (Ala: α-helix, Val: β-sheet, and Gly: β-turn or coil). Both hydrophobic amino acid (Val) and hydrophilic amino acid (Asp) are also opportunely included in the four amino acids encoded by the GNC code (Table 1), and are necessary for the folding of polypeptide chains into stable globular structures in water. Moreover, Asp with a functional group (carboxyl group), which is indispensable in the construction of a catalytic center on primeval proteins, is among in the four amino acids (Table 1).

At present, it is widely believed that proteins cannot be considered as the first materials in the creation of life because they cannot be replicated, even though [GADV]-amino acids should be more easily synthesized under prebiotic conditions than nucleotides in RNA and DNA. However, we noticed that [GADV]-proteins must synthesize similar but not identical [GADV]-proteins even in the absence of genes if the GNC primeval genetic code hypothesis is correct and if [GADV]-proteins have a catalytic function for peptide bond formation. The reason for this is because only four kinds of amino acids are used in [GADV]-proteins.

Table 1. Properties of [GADV]-Amino Acids[16].

Amino Acid	Hydropathy	α-Helix	β-Sheet	β-Turn	Group
Glycine	1.0	0.56	0.92	**1.64**	—
Alanine	1.6	**1.29**	0.90	0.78	—
Aspartic Acid	**–9.2**	1.04	0.72	**1.41**	**-COOH**
Valine	**2.6**	0.91	**1.49**	0.47	—

Bold letters indicate characteristic values of the amino acids for formation of secondary and tertiary structures of water-soluble globular proteins.

Of course, sequence diversity of [GADV]-proteins with 100 residues is still as high as about 10^{60}. However, it is assumed that every sequence composed of four amino acids would probably be detected at a probability of at least one time in a protein with only $4^4=256$ residues. Therefore, it can easily be imagined that middle-sized [GADV]-proteins composed of 256 amino acids are similar to each other.[5] In addition, the surface properties of [GADV]-proteins should be similar to each other, since there is a high probability that [D] and/or [G] with hydrophilic side chains should be located on surface regions and that [V] and/or [A] with hydrophobic side chains should be located in the inner cores of globular proteins in water. By taking these points into consideration, we reached the [GADV]-protein world hypothesis for the origin of life, which is based on pseudo-replication in the absence of genes (Fig. 6).[4,5]

In contrast, as described in the first section of this review, the RNA world hypothesis has been provided, suggesting RNA was amplified by self-replication and increased the diversity in the RNA world on the primitive earth based on the fact that RNA carries not only genetic information but also catalytic functions (Fig. 1)[12,13] and that the hypothesis is supported by many people. However, we have noticed that there are many weaknesses in the "RNA world hypothesis," which would be probably impossible to solve, as described below.[4,5]

(i) Nucleotides in RNA are organic compounds, which are far more complex, than are [GADV]-amino acids, as can be easily understood from comparing the numbers of atoms in the mononucleotides, UMP, CMP, AMP, and GMP (34~38), with those in [GADV]-amino acids (10~19). Thus, it would be much more difficult to synthesize the four nucleotides under prebiotic conditions than to synthesize the four amino acids. Indeed, no report indicating that nucleotides were synthesized from simple inorganic compounds under prebiotic conditions has yet been published.

(ii) Nevertheless, assume that nucleotides could be synthesized under prebiotic conditions, otherwise the RNA world would not be a reality. The number of hydroxyl groups on chiral carbon atoms in the nucleotides is so great that RNA cannot be synthesized by joining them with phosphodiester bonds in the absence of effective proteinous catalysts. On the contrary, peptide synthesis between carboxyl and amino groups on two amino acids should be much more easily accomplished.

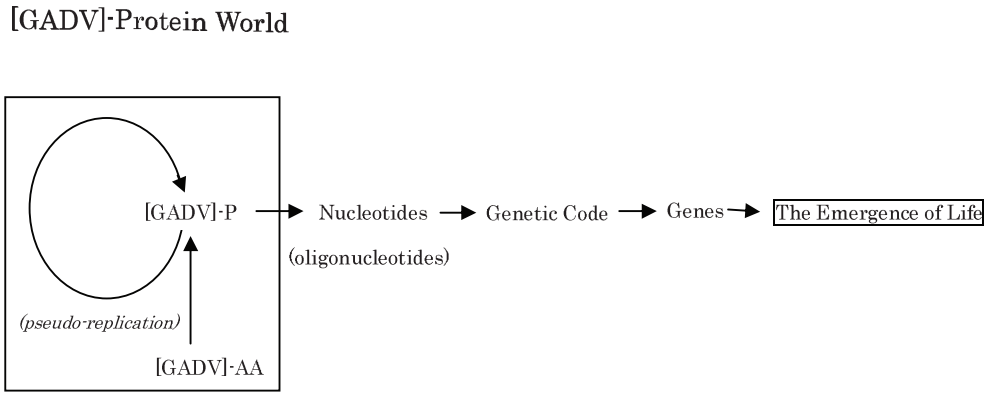

Fig. 6. [GADV]-protein world hypothesis for the origin of life, which we have postulated. The hypothesis states that life originated from the [GADV]-protein world, where [GADV]-proteins were amplified by pseudo-replication in the absence of any genetic function. The simple amino acid composition of [GADV]-proteins could enable pseudo-replication of the most primitive proteins. The [GADV]-proteins thus accumulated should produce nucleotides and oligonucleotides, leading to the formation of genetic code and genes. As a consequence, the first life emerged on the primitive earth.

(iii) Nevertheless, assume that RNA could be synthesized under prebiotic conditions, otherwise the RNA world would never be formed. However, it would be impossible for RNA to self-replicate because RNA without a stable tertiary structure is required to exhibit a genetic function on the nucleotide sequence as a template, whereas RNA must be folded into a stable tertiary structure to exhibit catalytic functions on RNA. This means that self-replication of RNA requires the above two self-contradictory structures. In fact, experimental results indicating that RNA molecules actually were self-replicated have not been previously reported even though much research on RNA self-replication has been performed.[13]

(iv) In addition to the above difficulties, there is another weakness in the RNA world hypothesis. There should be no relationship between the ability of RNA to self-replicate and the genetic function of RNA sequences for the synthesis of a protein. Therefore, even though RNA was self-replicated on the primitive earth, it is difficult to see how self-replicated RNA could acquire any genetic information for protein synthesis.

Thus, we consider that the RNA world never existed on the primitive earth.[4,5] Instead, we believe that the [GADV]-protein world hypothesis described above could surmount the weakness regarding protein replication by the introduction of a novel concept: the pseudo-replication of [GADV]-proteins in the absence of genes.[4,5]

Although many researchers have discussed the origins of genes, the genetic code, proteins, and life, they have treated the subjects of the fundamental system of life rather independently. Besides, we have now recognized that there are major weaknesses in their hypotheses for the origins of fundamental systems of life.[5] On the contrary, it could be possible to explain the four origins (genes, the genetic code, proteins, and life) comprehensively from the standpoint of the GNC–SNS primitive genetic code hypothesis, as described above (Fig. 3).[5] Thus, we firmly believe that the [GADV]-protein world hypothesis for the origin of life is a far more reasonable explanation of the emergence of life than is the RNA world hypothesis, and that the [GADV]-protein world could lead to formation of "RNA-protein world" and to the emergence of life.

Possible Steps to the Emergence of Life

We would now like to explain the possible steps to the emergence of life, based on the [GADV]-protein world hypothesis, as follows (Fig. 7).

(i) Prebiotic Synthesis of [GADV]-Amino Acids

[GADV]-amino acids could have been accumulated through a process of chemical evolution from simple inorganic compounds on the primitive earth. It is acknowledged by almost all researchers participating in the study of the origin of life that [GADV]-amino acids were rather easily synthesized on the primitive earth.

(ii) Formation of the [GADV]-Protein World

[GADV]-proteins could have been produced by random synthesis of peptide bonds among the four amino acids through

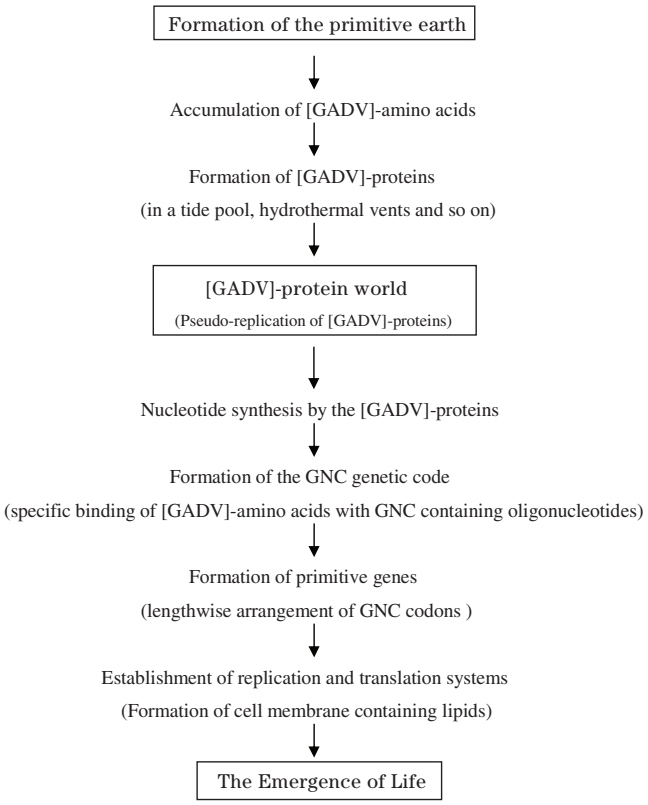

Fig. 7. Possible steps to the emergence of life explained from the perspective of the [GADV]-protein world hypothesis, which we have proposed. The hypothesis states that life originated from the [GADV]-protein world composed of [GADV]-proteins, which were amplified by pseudo-replication in the absence of genetic function.

repeated heat-drying processes in tide pools and/or polymerization of amino acids in thermal vents under the sea on the primitive earth. The [GADV]-protein world would thus be formed by the accumulated [GADV]-proteins. As a next step, the [GADV]-protein world was expanded and established through pseudo-replication of [GADV]-proteins in the absence of any genetic function.

This action might make it possible to produce microgranules or primeval cells surrounded by [GADV]-proteins through the side-chain of a hydrophobic amino acid, valine. In turn, this could accelerate the synthesis and accumulation of [GADV]-proteins more efficiently than before.

(iii) Formation of a Primeval Metabolic System by [GADV]-Proteins

The primeval metabolic system was successively formed by [GADV]-proteins with various amino acid sequences. In addition to the four amino acids, nucleotides and oligonucleotides (nucleic acid or RNA) were synthesized and accumulated by catalytic functions of the versatile [GADV]-proteins. This enabled the accumulation of various compounds required to form the most primitive and fundamental life system.

(iv) Establishment of GNC-Primeval Genetic Code and Formation of $(GNC)_n$ Primeval Genes

At the next step, GNC-primeval genetic code was established possibly through specific interaction between [GADV]-amino acids and oligonucleotides containing GNC (GNC primeval tRNAs).[4,5] A lengthwise arrangement of [GADV]-amino acid and GNC primeval tRNA complexes could accelerate the synthesis of [GADV]-proteins much more efficiently through the GNC primeval genetic code system.

It is assumed that a primitive translation system should have been formed through the creation of primitive ribosomal particles and that $(GNC)_n$ primeval genes were produced through phosphodiester bond formation among GNC codons.

(v) Formation of Double-Stranded RNA and the Emergence of Life

Single-stranded $(GNC)_n$ primeval genes were developed to produce double-stranded RNA. Thus, the basic life system was established by primitive double-stranded RNA genes. The appearance of double-stranded RNA enabled the inheritance and evolution of genetic information during the propagation of RNA sequences from ancestors to descendants. At this step, it is thought that the most primitive forms of "life" might have emerged on the primitive earth.

(vi) Development of $(SNS)_n$ Primitive Gene-SNS Primitive Genetic Code System

$(SNS)_n$ primitive genes and the SNS primitive genetic code were created through the evolution of $(GNC)_n$ primeval genes and the GNC primeval genetic code. It is likely that the creation of the $(SNS)_n$ genes and the SNS code was rather easily accomplished by coevolution between genes and genetic code, although it might have taken a very long time (Fig. 4).

Evidence for the [GADV]-Protein World Hypothesis

The most important factor in obtaining evidence for the [GADV]-protein world hypothesis is to confirm whether [GADV]-proteins can be pseudo-replicated to produce similar [GADV]-proteins. However, it is difficult to detect peptide synthetic activity of [GADV]-proteins because peptide synthesis is a thermodynamically unfavorable reaction. However,

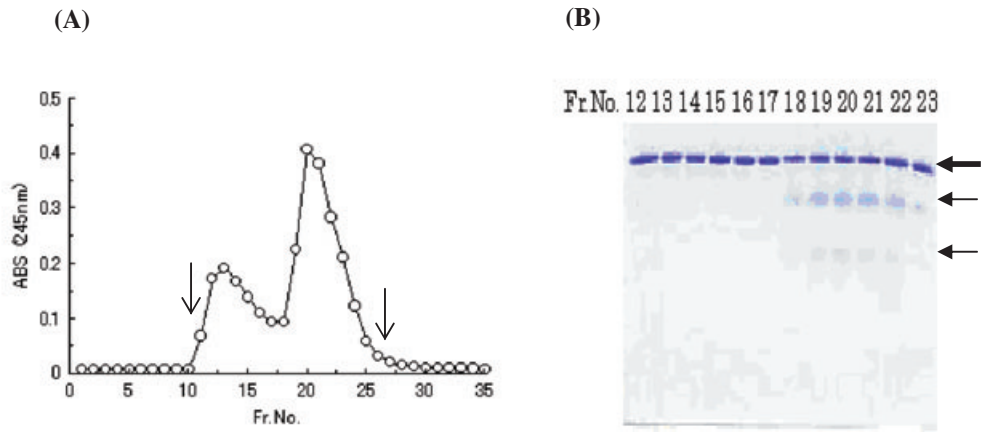

Fig. 8. (A) Sephadex G25 gel filtration chromatography of [GADV]-peptides, which was produced by 30 repetitions of heat-drying ([GADV]-P_{30}). The left and right fine arrows indicate the elution positions of the void volume and column volume estimated with blue dextran and riboflavin, respectively. (B) Hydrolytic activity of [GADV]-P_{30} toward peptide bonds in a natural protein, bovine serum albumin (BSA). Mixtures of BSA and the peptides in fractions from the column were incubated at 37°C for 6 days. The numbers written at the top of the SDS-PAGE electropherogram indicate the fraction number of the gel chromatography. The bold arrow and fine arrows indicate the positions of the native protein and hydrolytic fragments of the protein, respectively.

substances that can catalyze the degradation of something could possibly synthesize it from the degradation products, since microreversibility always exists in catalytic reactions including proteinous enzymes. Therefore, [GADV]-proteins should be able to be synthesized from their monomeric units, amino acids, if [GADV]-proteins could catalyze the hydrolysis of peptide bonds in proteins. Thus, in order to gather evidence on whether [GADV]-proteins can catalyze peptide bond formation among amino acids, we tried to detect whether [GADV]-proteins can hydrolyze peptide bonds in a natural protein, bovine serum albumin (BSA).[23]

Catalytic Activities of [GADV]-Peptides Produced by Pepeated Heat-Drying Processes

Several papers have previously reported the formation of peptides by heat-drying treatments and by a flow reactor simulating deep-sea hydrothermal vents, and so on.[21,24–28] However, no report has been published about catalytic activities against natural peptide bonds by the peptides. Alternatively, we tried to synthesize [GADV]-proteins from [GADV]-amino acids by repeated heat-drying treatments.[23] From the detection of peptide bond formation, it was confirmed that [GADV]-peptides were definitely produced by repeated heat-drying cycles of a [GADV]-amino acid aqueous solution containing $CuCl_2$ and NaCl.[26]

Thus, we first examined whether the [GADV]-P_{30}, which was obtained by heat-drying treatments repeated 30 times, has catalytic functions that hydrolyze some chemical bonds, such as the β-galactoside bond in 4-methylumbelliferyl-β-D-galactoside and the amide bond (peptide bond) in glycine-*p*-nitroanilide. The results revealed that [GADV]-P_{30} could hydrolyze the chemical bonds, yielding umbelliferone and *p*-nitroaniline, respectively (data not shown). [23]

Because it was expected that [GADV]-peptides would have a low catalytic activity, Sephadex gel filtration of the peptides was performed to remove possible contaminants of proteases and low molecular weight chemical compounds. The treatments were also effective in confirming the reproducibility of the probable low activity of the peptides and to estimate the apparent molecular weight of [GADV]-P_{30}. Two peaks were observed when [GADV]-P_{30} was subjected to the Sephadex G15 column (Fig. 8(A)). This means that several molecules of the peptides aggregated in the aqueous solution.[23]

In addition, to confirm the catalytic activity of [GADV]-P_{30} hydrolyzing peptide bonds in a protein, a mixture of BSA and [GADV]-P_{30} was held at 37°C for several days. The results, which were analyzed with SDS-polyacrylamide gel electrophoresis after holding the mixture for 6 days, are given in Fig. 8(B). Figure 8 clearly shows that [GADV]-P_{30} in an unaggregated form possesses hydrolytic activity against the peptide bonds in the natural protein, BSA.[23]

However, the color of the [GADV]-amino acid solution gradually changed from faint blue to yellowish green and a fluorescence emission from the sample developed as the number of heat-drying cycles increased. This indicated that fluorochromes, such as cyclized molecules, diketopiperazines, were formed in the solution in addition to [GADV]-peptides or -proteins.[23] This formation indicates that [GADV]-proteins, which were formed from [GADV]-amino acids in the absence

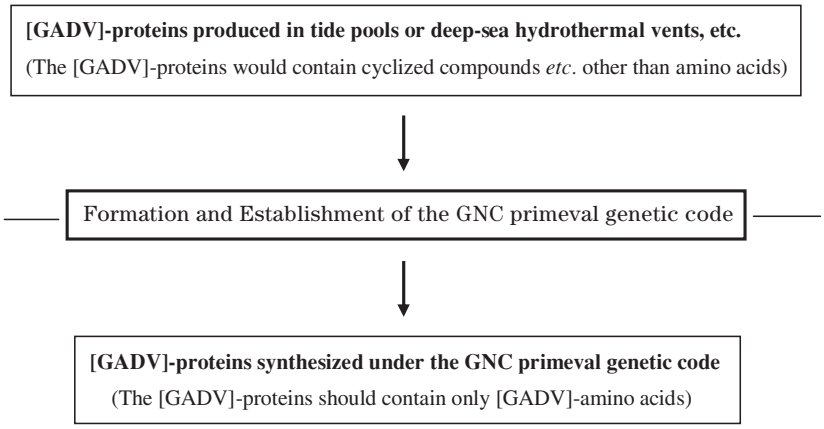

Fig. 9. [GADV]-proteins, which were produced before and after formation of the GNC primeval genetic code, must have had different chemical compositions and roles in the creation of the first life form (the emergence of life). The former proteins might have contained substances other than [GADV]-amino acids, such as cyclized compounds and have branched structures at aspartic acid residues in the molecules. In contrast, the latter proteins should have contained only [GADV]-amino acids linked by peptide bonds without branched structures. It is also thought that the former proteins played a role in the formation of the [GADV]-protein world, whereas the latter proteins played a role in the establishment and development of the [GADV]-protein world and in the formation of the most primitive genes leading to the emergence of life.

of a genetic code and genes, must contain cyclized compounds as well.

In contrast, it is supposed that [GADV]-proteins synthesized after the establishment of the GNC genetic code, which encodes [GADV]-amino acids, must be composed of only amino acids because the genetic code would provide a peptide synthetic pathway of [GADV]-amino acids (Fig. 9).[23] Therefore, it is important to confirm whether both [GADV]-proteins produced by repeated dry-heating processes, which would play an important role in the formation of the [GADV]-protein world, and those obtained by solid phase synthesis, which would play a role in the establishment of the world, have hydrolytic activities toward peptide bonds. The two types of [GADV]-proteins also correspond to the [GADV]-proteins synthesized before and after formation of the GNC-primeval genetic code, respectively (Fig. 9).[23]

Properties of [GADV]-Random Octapeptides Synthesized with Peptide Synthesizer

[GADV]-octapeptides with random sequences were synthesized with a peptide synthesizer, corresponding to [GADV]-peptides synthesized under the most basic genetic code, GNC (Fig. 3).[23] In contrast to [GADV]-P_{30}, the [GADV]-peptides synthesized with the peptide synthesizer did not contain any fluorochromophore, as was expected.

The [GADV]-random octapeptides were dissociated by the addition of 4 M urea, as seen in Fig. 10(A).[23] This indicates that several molecules of octapeptides made aggregates in an aqueous solution probably through hydrophobic interaction among the side chain of valine in the peptides. We thus examined whether the [GADV]-octapeptides can catalyze the hydrolysis of peptide bonds in BSA. As shown in Fig. 10(B), the [GADV]-octapeptides did catalyze the hydrolysis of the peptide bonds,[23] but the peptide bonds were not substantially hydrolyzed in the absence of the octapeptides (data not shown), suggesting that the peptides could form peptide bonds using surrounding [GADV]-amino acids and would effectively accumulate [GADV]-peptides or [GADV]-proteins through the microreversibility of catalysts. Based on the facts, it can be concluded that even [GADV]-peptides without cyclized compounds such as diketopiperazines could catalyze the synthesis of peptide bonds.

The high hydrolytic activity of the random octapeptides suggests that the [GADV]-proteins produced under the GNC primeval genetic code would contribute to the establishment of the [GADV]-protein world and to the acceleration of the steps to the emergence of life (Fig. 7). Effective accumulation of the [GADV]-proteins by pseudo-replication might also accelerate nucleotide formation through their high catalytic activities, which could lead to formation of an "RNA-protein world" following the [GADV]-protein world.[23]

Conclusion

Even if [GADV]-proteins in a [GADV]-protein world could accumulate [GADV]-proteins through their pseudo-

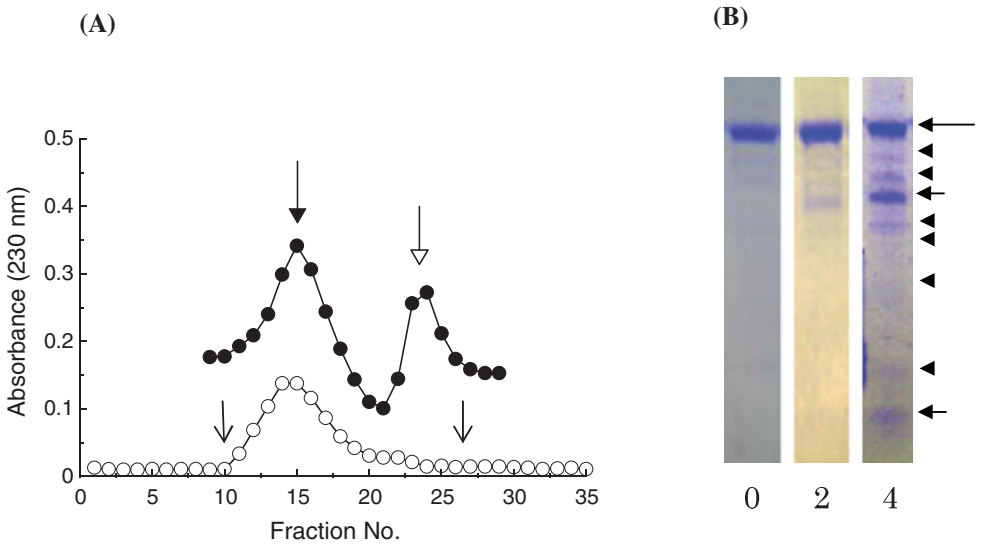

Fig. 10. (A) Sephadex G25 gel filtration chromatography of [GADV]-random octapeptides. The open and closed circles indicate absorbance at 230 nm of the octapeptides in the absence and presence of 4 M urea, respectively. The arrows with closed and open heads indicate the elution positions of aggregated and dissociated forms of the random octapeptides; the left and right fine arrows indicate void volume and column volume, respectively. (B) Hydrolytic activity of [GADV]-octapeptides, which was produced by a protein synthesizer, toward peptide bonds in BSA. A mixture of BSA and the peptides in a peak fraction (fraction number 14) from the Sephadex G25 chromatography in the absence of urea was incubated at 37°C for 0, 2, and 4 days. The numbers written at the bottom of the electropherograms indicate the number of days the mixture was held at 37°C. The long arrow, short arrows, and arrowheads indicate the band positions of native BSA and major and minor fragments of the protein, respectively. In the control experiments carried out in the absence of the octapeptides, the peptide bonds in BSA were not substantially hydrolyzed (data not shown).

replication in the absence of genes,[4,5] these proteins were not even the simplest of life forms, since the proteins could not evolve in the protein world while genes were absent. It is important, however, to note that the [GADV]-proteins should be indispensable chemical components in leading to the emergence of life and that life did not originate from the RNA world. The reason for this is that it would have been very difficult to produce nucleotides as components of RNA at a meaningful concentration and to replicate RNA on the primitive earth without the protein world.[29,30]

In contrast, based on the perspective of the [GADV]-protein world hypothesis, it can be reasonably assumed that [GADV]-proteins produced in the [GADV]-protein world would accumulate nucleotides through their high catalytic activities and create the most primitive genetic code, GNC (Fig. 7). It is also assumed that the most primitive genes could be created by lengthwise arrangement of the GNC genetic code. Although sufficient evidence for the steps to the emergence of life have not yet been obtained, the [GADV]-protein world hypothesis is superior to the RNA world theory because reasonable steps to the emergence of life could be assumed as a path from simple to complex chemical compounds (Fig. 7). The experimental results also support the [GADV]-protein world hypothesis for the origin of life (Figs. 9 and 10). Thus, we conclude that life must have originated from the [GADV]-protein world on through the RNA-protein world (Fig. 7).

REFERENCES

[1] Woese, C. R.; Fox, G. E. J Mol Evol 1977, 10, 1.
[2] Lazcano, A.; Forterre, P. J Mol Evol 1999, 49, 411.
[3] Barbenko, V. N.; Krylov, D. M. Nucl Acids Res 2004, 32, 5029.
[4] Ikehara, K. Seibutsukagaku (in Japanese) 1999, 51, 43.
[5] Ikehara, K. J Biosci 2002, 27, 165.
[6] Pross, A. Orig Life Evol Biosph 2004, 34, 307.
[7] Oparin, A. I. The Origin of Life on Earth; Oliver and Boyd: Edinburgh; 1957.
[8] Wachtershauser, G. J. Theor Biol 1997, 187, 483.
[9] Melosh, H. J. Nature 1988, 332, 687.
[10] Kruger, K.; Grabowski, P. J.; Xaug, A. J.; Sands, J.; Gottschling, D. E.; Cech, T. R. Cell 1982, 31, 147.
[11] Guerrier-Takada, C.; Gardiner, K.; Marsh, T.; Pace, N.; Altman, S. Cell 1983, 35, 849.
[12] Gilbert, W. Nature 1986, 319, 618.

[13] Gesteland, R. F.; Cech, T. R.; Atkins, J. F. The RNA World; Cold Spring Harbor Laboratory Press; 1999.
[14] Ikehara, K.; Amada, F.; Yoshida, S.; Mikata, Y.; Tanaka, A. Nucl Acids Res 1996, 24, 4249.
[15] Ikehara, K. Seibutsukagaku (in Japanese) 1998, 50, 44.
[16] Ikehara, K.; Yoshida, S. Viva Origino 1998, 26, 301.
[17] Ikehara, K. Viva Origino 1998, 26, 311.
[18] Ikehara, K; Omori, Y.; Arai, R.; Hirose, A. J Mol Evol 2002, 54, 530.
[19] Berg, J. M.; Tymoczko, J. L.; Stryer L. Biochemistry; W. H. Freeman and Company: New York; 2001, p 67, p 334.
[20] Ikehara, K.; Okazawa, E. Nucl Acids Res 1993, 21, 2193.
[21] Miller, S. L; Orgel, L. E. The Origin of Life on the Earth; Prentice-Hall, Inc: Englewood Cliffs; 1974.
[22] Trifonov, E. N. J Biomol Struc Dyn 2004, 22, 1.
[23] Ohba, T.; Fukushima, J.; Maruyama, M.; Iwamoto, R.; Ikehara, K. Orig Life Evol Biosph 2005, in press.
[24] Yanagawa, H.; Makino, Y.; Sato, K.; Nishizawa, M.; and Egami, F. Orig Life 1984, 14, 163.
[25] Sakurai, M.; Yanagawa, H. Orig Life 1884, 14, 171.
[26] Rode, B. M.; Eder, A. H.; Yongyai, Y. Inorg Chim Acta 1997, 254, 309.
[27] Imai, E.; Honda, H.; Hatori, K.; Matsuno, K. Orig Life Evol Biosph 1999, 29, 249.
[28] Imai, E.; Honda, H.; Hatori, K.; Brack, A.; Matsuno, K. Science 1999, 283, 831.
[29] Shapiro, R. Orig Life 1984, 14, 565.
[30] Shapiro, R. Orig Life Evol Biosph 1988, 18, 71.