Article

Not peer-reviewed version

Evolutionary Pathway of the Genetic Code deduced based on Coevolution Theory

Kenji Ikehara *

Posted Date: 20 February 2025

doi: 10.20944/preprints202502.1594.v1

Keywords: Origin of genetic code; Evolution of genetic code; GNC genetic code; SNS genetic code; RNY genetic code; Coevolution theory; Evolution of metabolic pathways

 doi:10.20944/preprints202502.1594.v1

Article

Evolutionary Pathway of the Genetic Code deduced based on Coevolution Theory

Kenji Ikehara [1,2]

[1] G&L Kyosei Institute in Keihanna Academy of Science and Culture (KASC), Keihanna interaction plaza, Lab. Wing 3F, 1-7 Hikaridai, Seika-cho, Souraku, Kyoto 619-0237, Japan; ikehara@cc.nara-wu.ac.jp; Tel.: +81-774-73-4478

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

Abstract: The coevolution theory suggests that origin and evolution of the genetic code proceed along with development of synthetic pathway of amino acids. The coevolution theory is undoubtedly a correct idea, because the genetic code could not evolve, if synthetic pathways of new amino acids were not invented and the new amino acids did not accumulate at a large amount in the cells. On the other hand, studies on metabolic pathways have progressed and information about metabolic pathways obtained thus far has been stored in KEGG PATHWAY Database. Then, amino acid metabolic pathways were analyzed in detail in order to answer to the following questions. (1) The genetic code originated from what type of genetic code? (2) The genetic code has reached to the universal genetic code, passing through what type of intermediate genetic code, SNS code or RNY code? From the results, it could be reconfirmed that the genetic code originated from the GNC primeval genetic code and evolved to the universal genetic code passing through SNS genetic code. Lastly, limits and future prospects of the coevolution theory are also discussed.

Keywords: Origin of genetic code; Evolution of genetic code; GNC genetic code; SNS genetic code; RNY genetic code; Coevolution theory; Evolution of metabolic pathways

1. Introduction

As a matter of course, studies on origin and evolution of the genetic code, which mediates between two core members (gene and protein) in the genetic system, are quite important to understand the origins and evolution of gene and protein. Genes and proteins use codons and amino acids specified by the genetic code, respectively. Therefore, origin of life could be solved through elucidation of origin of the genetic system composed of gene, genetic code and protein (Figure 1).

On the other hand, metabolic pathways are driven by proteins or enzymes, which are produced by expression of genetic information written into genes. The enzymes are synthesized with amino acids, which are produced through metabolic pathways. Therefore, amino acids synthesized through the most primitive metabolic pathways should be used in the first genetic code. Inversely stating this, amino acids used in the first genetic code could be determined by knowing the amino acids, which were produced through the first amino acid synthetic pathways (Figure 1).

Thus, elucidation of origin and evolution of the genetic code is a quite important matter, which may lead to solving not only origins and evolution of gene and protein but also origins of evolution of metabolism and life. Nevertheless, the origin and evolutionary process of the genetic code, especially, the evolutionary process, about what genetic code was used as an intermediate code bridging over the first genetic code with the universal genetic code has not been well made clear. On the other hand, it is well known that coevolution theory suggests that the genetic code and the biosynthetic relationships between amino acids evolved in parallel. Then, in this article, the origin and evolution of the genetic code are discussed from a viewpoint of the coevolution theory [1–5], especially of evolution of amino acid synthetic pathways.

 doi:10.20944/preprints202502.1594.v1

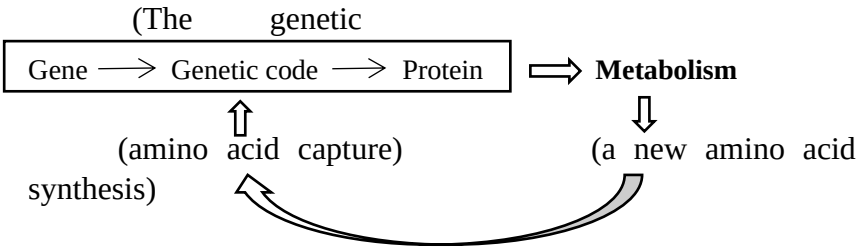

Figure 1. Metabolic pathways are driven by proteins or enzymes (white bold arrow). Proteins are produced through the genetic system mainly composed of three members, gene, genetic code and protein (thin black arrows). The genetic code has evolved by incorporation of new amino acids (upward white bold arrow), which were synthesized through metabolic pathways (downward white bold arrow and curved bold arrow).

2. Method for Analysis of Metabolic Pathway

Metabolic pathways necessary to analyze amino acid syntheses were extracted from KEGG PATHWAY Database (1. Metabolism, Amino acid) [6]. A precursor-product relationships and chemical structures of amino acids, intermediate metabolites etc. were analyzed.

3. Why genetic code and amino acid synthetic pathway coevolve?

The coevolution theory suggests that the genetic code coevolved with invention of biosynthetic pathways for new amino acids [1–5]. That is the reason why the organization of the genetic code is determined by relationships of precursor amino acid-product amino acid [1–5]. Of course, an amino acid produced upon formation of a new synthetic pathway should trigger formation of a new genetic code on the way of evolution. Therefore, the idea would be naturally and always valid, because evolution of the genetic code encoding amino acids is determined by the order of amino acids, which were produced through metabolic pathways and accumulated at a large amount in cells.

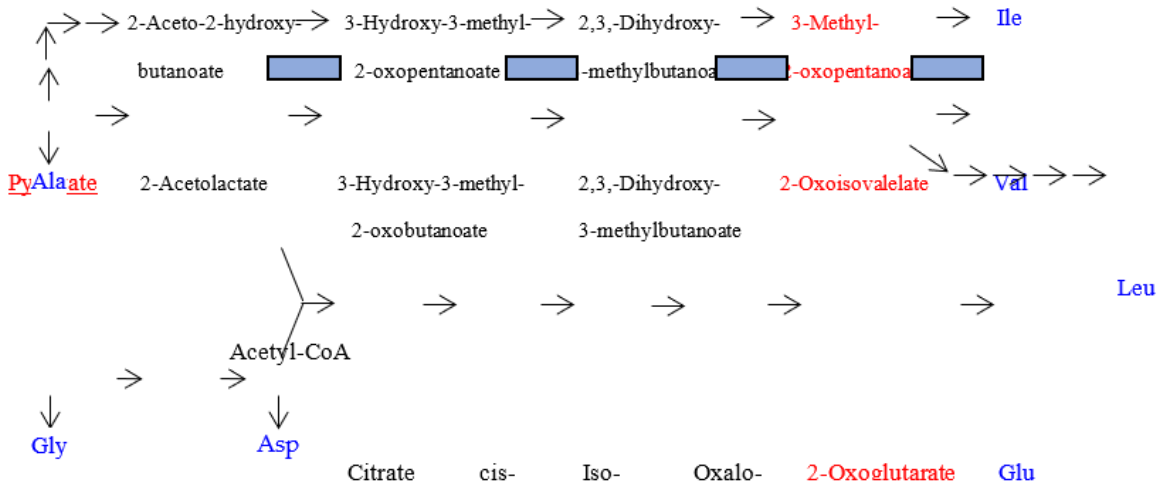

Figure 2. Synthetic pathways of five amino acids (Gly, Ala, Asp, Val and Glu) encoded by G-start codons and of two hydrophobic amino acids, Ile and Leu. The accumulation order of the seven amino acids was deduced based on the number of reaction steps from glyoxylate or pyruvate (indicated by underlined red letters) to the respective amino acids, as (Gly, Ala), Asp, (Val, Glu) and (Leu, Ile). Ketoacids used direct amination are written by red letters. One reaction step from 2-Isopropyl-3-oxosuccinate to 4-Methyl-2-oxopentanoate is omitted from the figure, because the step proceeds spontaneously. Blue boxes indicate the same enzymes used for both Val and Ile syntheses.

 doi:10.20944/preprints202502.1594.v1

Note that similar things should be observed in the cases of formation of, not only amino acid synthetic pathway but also any metabolite synthetic pathway, because the idea of the coevolution theory could be applied to studies on evolutionary process of the metabolic pathways [6]. In some cases, a new metabolic pathway is formed by using of an intermediate of previously existed metabolic pathway. For example, Leu synthetic pathway is formed by using an intermediate, 2-oxoisovalerate, for Val synthesis as a starting molecule (Figure 2) [6]. In this case, it would be obvious that Leu synthetic pathway was formed after Val synthetic pathway was completed. Thus, it would be easily understood that the coevolution theory is a valid idea, because evolution of genetic code should be triggered by accumulation of an amino acid, which was produced through a newly formed metabolic pathway. Note that it can be also applied in the cases when a new synthetic pathway is formed by using enzymes driving a part of a previously existed metabolic pathway. There is an example that Ile synthetic pathway is formed by connecting a similar but different chemical compound (2-oxo-2-hydroxy butanoate) produced through a new metabolic pathway from pyruvate with the enzyme system starting from 2-acetolactate (Figure 2) [6]. In this case too, it can be concluded that Ile synthetic pathway was formed after completion of Val synthetic pathway (Figure 2).

Thus, it can be reconfirmed that the coevolution theory is a valid idea based on some applicable examples above, in addition to the results obtained thus far [1–5]. In the next Section 4, some conditions, which must be satisfied when the coevolution theory is applied to evolutionary process of the genetic code, are discussed.

4. Conditions making it possible to study on the origin and evolutionary process of genetic code

It is also important to confirm that there exist some problems before evolutionary process of the genetic code is discussed based on the coevolution.

4.1. Formation of a new amino acid synthetic pathway leads to evolution of the genetic code

It is definite that synthesis of a new amino acid advanced evolutionary process of the genetic code as expected by the coevolution theory [1–5]. Therefore, it would be possible to deduce the origin and evolutionary process of the genetic code, if it could be understood what amino acid was synthesized through a new metabolic pathway and accumulated in cell structure.

4.1.1. The order of amino acid capture into genetic code could be determined from analysis of modern metabolic map

It would become possible to understand the origin and evolutionary process of the genetic code from analysis of modern amino acid metabolic pathways, if the ancient pathways have continued to be used still now.

(1) It is generally quite difficult to change from one metabolic pathway for synthesis of an amino acid to another new metabolic pathway for the same amino acid synthesis. The reason, why it is difficult to change from a metabolic pathway established previously to another new metabolic pathway, is because enzymatic activities of new unrefined and immature proteins, which were newly synthesized and were used in the new metabolic pathway, should be quite low and an assembly of the new metabolic pathway should be also fragile, although in some cases a new product synthesized though a new metabolic pathway might complement quantitatively a deficiency of the same product, which was produced through previously existed metabolic pathway.

(2) It is also essentially impossible to change an amino acid assigned to a codon to another new amino acid, because it is necessary to once change from the assigned codon to unassigned codon, which does not specify any amino acid. However, the change generates multiple termination codons to cause lethal mutations. That is the reason, why genetic code, which was once established, can be generally unchanged.

 doi:10.20944/preprints202502.1594.v1

As described above, both GNC code and [GADV]-amino acid synthetic pathways should remain as vestiges in modern universal genetic code and metabolic pathways, which modern organisms use in these days. Therefore, what genetic code and metabolic pathways were used in lives on primitive Earth could be revealed by analyzing universal genetic code and modern metabolic pathways for amino acid syntheses.

4.1.2. Origin of the genetic code deduced from analyses of universal genetic code and modern metabolic pathways

Confirm here points to attention when evolutionary pathway of the genetic code is deduced by analyses of amino acid metabolic pathways or the coevolution theory.

(1) It must be first noted that an amino acid, which is produced upon formation of a new synthetic pathway, did not always accumulate at a large amount in cells, because accumulation amount of the amino acid should become small, if the amino acid was used as substrates for syntheses of some other organic compounds. For example, Asp, which is synthesized by a direct amination of oxaloacetate, is used for not only syntheses of three amino acids, Thr, Asn and Lys, but also formation of aromatic rings of purine and pyrimidine nucleotides (Figure 3) [6]. Therefore, Asp should not accumulate immediately at a large amount in cells. On the contrary, another acidic amino acid, Glu, which is also synthesized by a direct amination of 2-oxoglutarate, could easily accumulate at a large amount in cells, because Glu was not used as a precursor molecule for syntheses of important chemical compounds like nucleotides except three amino acids, Pro, Asn and Arg (Figure 3) [6]. That is, it is essential for an amino acid to be used in the genetic code that the amino acid actually accumulates at a large amount in cells.

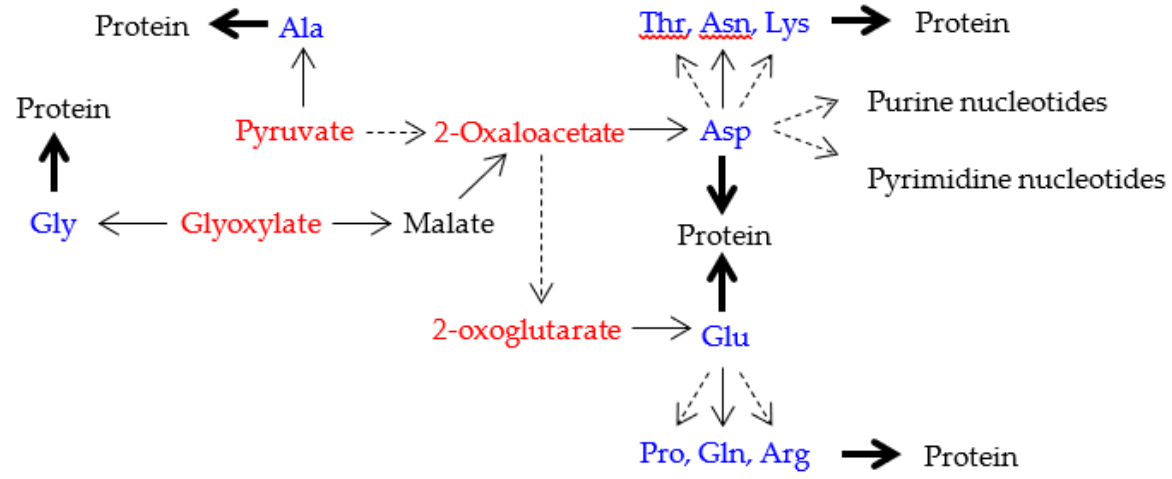

Figure 3. Metabolic pathways starting from four ketoacids (written with red letters). Asp is used as precursor molecules for purine and pyrimidine nucleotides. Single-step reactions and multi-step reactions are indicated by thin arrows and broken arrows, respectively. Bold black arrows indicate that amino acids are used for protein synthesis.

(2) Inversely, an amino acid, which was synthesized and actually accumulated through a new metabolic pathway at a sufficiently large amount in cells, is not always captured into the genetic code. The reason is because it is one of conditions whether or not the amino acid, which was produced through a new metabolic pathway, could contribute to functional enhancement of proteins, which were synthesized under the new genetic code. For example, as well known, 2-aminobutylate (2-AB), which should be easily synthesized and accumulate at a large amount in cells, was not incorporated into the universal genetic code. The reason is because 2-AB (a weak hydrophobic and α-helix forming amino acid) was competed with Ala (a weak hydrophobic and α-helix forming amino acid) and more simply structured Ala than 2-AB was adopted as a natural amino acid.

 doi:10.20944/preprints202502.1594.v1

Therefore, necessary conditions for a new amino acid to be incorporated into previously existing genetic code are as follows.

(1) Accumulation of an amino acid, which is synthesized through a new metabolic pathway, in cells.

(2) Functional enhancement of proteins synthesized under a new genetic code, which captured an amino acid.

That is, formation of a new metabolic pathway for an amino acid is only a necessary condition but not a sufficient condition for incorporation of the amino acid into a new genetic code.

5. The origin of the genetic code deduced from coevolution theory

In this article, the following three matters are mainly discussed based on the coevolution theory. (1) What was the first genetic code? (2) How has the first genetic code evolved to the universal genetic code via what type of an intermediate code? (3) Why must the genetic code pass through the intermediate code?

The coevolution theory, which advocates that genetic code has evolved together with development of amino acid metaboism (Figure 1), is one of important concepts for explaining evolutionary process of the genetic code in relation to amino acid synthetic pathways [1–5]. The coevolution theory is also a reasonable idea, which can be applied to the origin of the genetic code.

Then, first reconfirm and discuss the origin of genetic code in relation to amino acid metabolism. However, it is necessary to know what was the first genetic code, in order to discuss on the origin and evolution of metabolism, because some primitive genetic code hypotheses have been proposed until now, for example, GC code hypothesis [7], the four column theory [8] (Table 1) and so on.

However, both ideas would be unreasonable for the first genetic code. The reasons are as follows

(1) GC code hypothesis (Table 1A)

Both hydrophobic amino acid and β-sheet forming-amino acid are not contained in the GC code. Instead, two turn/coil-forming amino acids, Pro and Gly, are contained in the four amino acids [7]. Therefore, any polypeptide chain synthesized under the GC code could not be folded into water-soluble globular structure, which is one of prerequisite conditions to express catalytic activity.

(2) Four olumn theory (Table 1B)

Similarly as GNC primeval genetic code [9,10], four [GADV]-amino acids are used in the genetic code, which is supposed in the four column theory [8]. However, it would be necessary in the four column theory to decode the genetic code as using wobble recognition at the first and the third codon positions. However, it would be impossible to form base pairs through the wobble recognition without any elaborate apparatus like ribosome in the first genetic code era, because base-pair formation should be carried out as seeking for the most stable base position. Otherwise, 64 tRNAs must be prepared for use of only four amino acids. In addition, it is necessary to once change from the assigned codon to unassigned codon on evolutionary process of the genetic code. However, the change would be impossible because of generation of multiple termination codons causing lethal mutations. Therefore it must be concluded that the first genetic code considered in the four column theory never be realized.

Table 1. Two genetic code tables, which have proposed as the first genetic code. **(A)** GC code hypothesis [7] and **(B)** Four column theory [8]. The GC code is highly GC-rich and the four column theory is composed of four [GADV]-amino acids. Both of which are partly similar to GNC primeval genetic code, which we have proposed [9,10].

 doi:10.20944/preprints202502.1594.v1

(A) GC code

	C	G	
C	Pro	Arg	C
	Pro	Arg	G
G	Ala	Gly	C
	Ala	Gly	G

(B) Four column

	U	C	A	G	
U	Val	Ala	Asp	Gly	U
C					C
A					A
G					G

These mean that it is not sufficiently understood still now even that the universal genetic code was originated from what type of genetic code. One of the reasons would be, probably because the origin of genetic code has been discussed not comprehensively but individually until now, irrespective of the fact that the genetic code plays an important role in connecting gene with protein (Figure 1).

On the other hand, we proposed GNC-SNS primitive genetic code hypothesis about 20 years ago [9,10]. According to the hypothesis, it is considered that the universal genetic code originated from GNC code composed of four GNC codons and four [GADV]-amino acids and evolved via SNS code composed of 16 codons and 10 amino acids. Then, newly reconfirm plausibility of the GNC primeval genetic code hypothesis.

(1) Protein containing [GADV]-amino acids at roughly equal amounts satisfies four conditions (hydrophobicity/hydrophilicity, α-helix, β-sheet, turn/coil formabilities) necessary to water-soluble globular protein formation [9,10]. This means that water-soluble but unrefined and immature [GADV]-proteins could be formed even by random joining of [GADV]-amino acids, because [GADV]-amino acid composition containing roughly equal amounts of [GADV]-amino acids is one of protein 0^{th}-order structures, which give a frame necessary to form water-soluble globular structure [10,11]. The characteristics of [GADV]-amino acids made it possible to select the four amino acids in the messy environments on the primitive Earth [12].

(2) Three anticodon stem-loops (AntiC-SLs) containing one of GNC triplets except GUC are stable without chemical modification in the loop [10,12]. In addition, the AntiC-SLs containing one of GNC triplets could make two stable dimers vertically bound through triplet base pairs (5′GGC3′/3′CCG5′ and 5′GUC3′/3′CAG5′) [13]. Thus, four GNC triples could be selected as anticodons for establishment of GNC primeval genetic code.

(3) A reasonable evolutionary process of tRNA could be deduced from analyses of anticodon stem base sequences of *Pseudomonas aeruginosa* tRNAs [10,14].

Evolutionary process of the genetic code could be deduced more correctly on the definite basis of the first genetic code, GNC, if it could be reconfirmed that the GNC primeval genetic code hypothesis is valid from a standpoint of the coevolution theory.

A hierarchy between amino acids encoded by codons specified by the first base and precursor molecules of the respective amino acid syntheses can be seen. For example, as seen in Figure 2, four amino acids encoded by G-start codons, Gly, Ala, Asp and Val, are synthesized by direct amination of the corresponding ketoacids, glyoxylate, pyruvate, oxaloacetate and oxoisovalerate (Figures 2 and 3) [6]. Therefore, it can be concluded that the four amino acids, Gly, Ala, Asp and Val, are the amino acids used in the first genetic code, because the four amino acids can be synthesized without using any other amino acid as a precursor molecule [6]. Val, which is produced at five reaction steps from pyruvate, is the most simply structured amino acid among natural hydrophobic amino acids (Figure 2 and Table 2). Furthermore, it can be understood that the synthetic pathway of Glu was formed after metabolic pathway for Asp synthesis was formed, because Glu is synthesized at four steps from oxaloacetate or a direct precursor molecule of Asp. Thus, the results obtained by analyses of amino acid metabolic pathways are consistent with the GNC primeval genetic code hypothesis insisting that the first genetic code was GNC encoding [GADV]-amino acids [14].

 doi:10.20944/preprints202502.1594.v1

Table 2. Relationship between (product) amino acid and precursor amino acid. **(A)** Four amino acids written in blue letters indicate [GADV]-amino acids encoded be G-start codons. Non-amino acid precursor molecules are written in parenthesis. The number written in parenthesis shows the number of reaction steps from respective precursor molecules. PRPP and Pen-p cycle means phosphoribosyl- pyrophospate and pentose-phosphate cycle, repressively. The number (4) described in double parentheses indicates reaction steps using the same enzymes with Val synthetic pathway (Figure 2).

(A)

Amino acid	Precursor amino acid	Amino acid	Precursor amino acid
G-start codons		Ala	(pyruvate) (1)
Gly	(glyoxylate) (1)	Val	(pyruvate) (5)
Asp	(glyoxylate) (3)	Glu	(oxaloacetate) (4)
C-start codons		His	(PRPP) (10)
Leu	(2-oxoisovalelate) (5)	Gln	Glu (1)
Pro	Glu (4)	Arg	Glu (8)
A-start codons		Asn	Asp (1)
Ile	(pyruvate) (5+(4))	Lys	Asp (7)
Thr	Asp (5)	Met	Cys (3)
U-start codons		Tyr	(Pen-p cycle) (10)
Phe	(Pent-p cycle) (10)	Cys	Ser (2)
Ser	Gly (1)	Trp	(chorismate) (5 or 6)

Note: Some characteristics can bee seen in the above table. (1) amino acid metabolic pathways of all the five amino acids (Gly, Ala, Asp, Val and Glu) using a G-start codon do not use any precursor amino acid. (2) Reaction steps from pyruvate more than 8 are used in the synthetic pathways of two hydrophobic amino acids (Leu and Ile) with a long side chain. (3) Many reaction steps (more than 10) counted from the initial precursor molecules are also used in the synthetic pathways of three aromatic amino acids (Phe, Tyr and Trp). (4) Intermediate molecule (2-oxoisovalerate) for Val synthesis is used in Leu synthetic pathway and a series of enzymes on the Val synthetic pathway are used for Ile synthesis. (5) PRPP is used at the initial step of His synthetic pathway. (6) Glu and Asp are used as a precursor amino acid in synthetic pathways for three amino acids encoded by C-start codon and A-start codon, respectively. Pent-p cycle is a abbreviation of pentose-phosphate cycle.

(B) The universal genetic code table summarized for each base at the first codon position. (Upper left) G-start codon table. (Upper right) C-start codon table. (Lower left) A-start codon table. (Lower right) U-start codon table. Term in the U-start codon table indicates termination codon.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 February 2025 doi:10.20944/preprints202502.1594.v1

	U	C	A	G	
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

	U	C	A	G	
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G

	U	C	A	G	
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Term	Term	A
	Leu	Ser	Term	Trp	G

Lei and Burton describes in their paper that the genetic code initially evolved to synthesize polyglycine as a cross-linking agent to stabilize protocells and, thereafter, the code sectored from a glycine code to a four amino acid code to an eight amino acid code to an ~16 amino acid code to the standard 20 amino acid code with stops [15]. Di Giulio also described that based on a result of the coevolution theory, the very earliest phases of genetic code the type GNN [4]. Thus, their ideas about the origin are consistent with GNC-SNS primitive genetic code hypothesis [9].

6. Evolutionary pathway of the genetic code viewed from coevolution theory or amino acid metabolism

Of course, a reliable genetic code connecting the first genetic code with modern universal genetic code could not be obtained, if the intermediate genetic code is discussed under a wrong idea about the first genetic code (Figure 4 (A)). However, it has been fortunately confirmed that the first genetic code was GNC code based on the results analyzed from a view point of the coevolution theory (Section 5).

(A)

The first genetic code (GNC primeval genetic code) ⟹ (an intermediate genetic code) S(C)NS code? or R(A)NY code? ⟹ Universal genetic code

(B)

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

(C)

	U	C	A	G	
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
G	**Val**	**Ala**	**Asp**	**Gly**	U
	Val	Ala	Asp	Gly	C

Figure 4. (A) Two evolutionary pathways of the genetic code, which passed through as an intermediate genetic code, **(B)** SNS code or **(C)** RNY code. It is considered that SNS code and RNY code were formed by piling up

 doi:10.20944/preprints202502.1594.v1

CNS code on GNS code and by piling up ANY code on GNY code, respectively. Amino acids written with bold letters indicate the four amino acids encoded by GNC primeval genetic code.

Then, discuss in this Section what was the intermediate genetic code connecting the GNC primeval genetic code with modern universal genetic code (Figure 4A). For the purpose, it is first described what type of intermediate genetic codes have been proposed. The first one is SNS code hypothesis (Figure 4B) [9] and the second one is RNY code hypothesis (Figure 4C) [16,17].

6.1. Analysis of Evolutionary Process of the Genetic Code

As described in Section 5, it has been reconfirmed from the viewpoint of the coevolution theory that the first genetic code was GNC code. Then, consider from viewpoint of the coevolution theory what type of genetic code was used to connect the GNC code with universal genetic code.

The results seen in Table 2, two cases of amino acids encoded by C- and A-start codons contrast with each other. Three (Pro, Gln, Arg) out of five amino acids encoded by C-start codons are synthesized as a starting molecule, Glu, one of five [GADVE]-amino acids, which are encoded by G-start codons (Figure 5) [6]. This means that metabolic pathways of the three amino acids were formed after metabolic pathway for Glu synthesis was established. On the other hand, Leu is synthesized with the last intermediate (2-oxoisovalerate) on Val synthetic pathway (Figure 2). This indicates that the synthetic pathway of Leu was formed after the metabolic pathway of Val was established. Note that His is synthesized at ten steps starting from the first reaction using PRPP and ATP (Figure 5) [6]. Therefore, it is considered based on the coevolution theory, that His is captured into a genetic code after formation of ATP synthetic pathway and probably establishment of GNS (GNC + GNG) code. Thus, it can be supposed that all the five amino acids encoded by C-start codons were incorporated into GNS code.

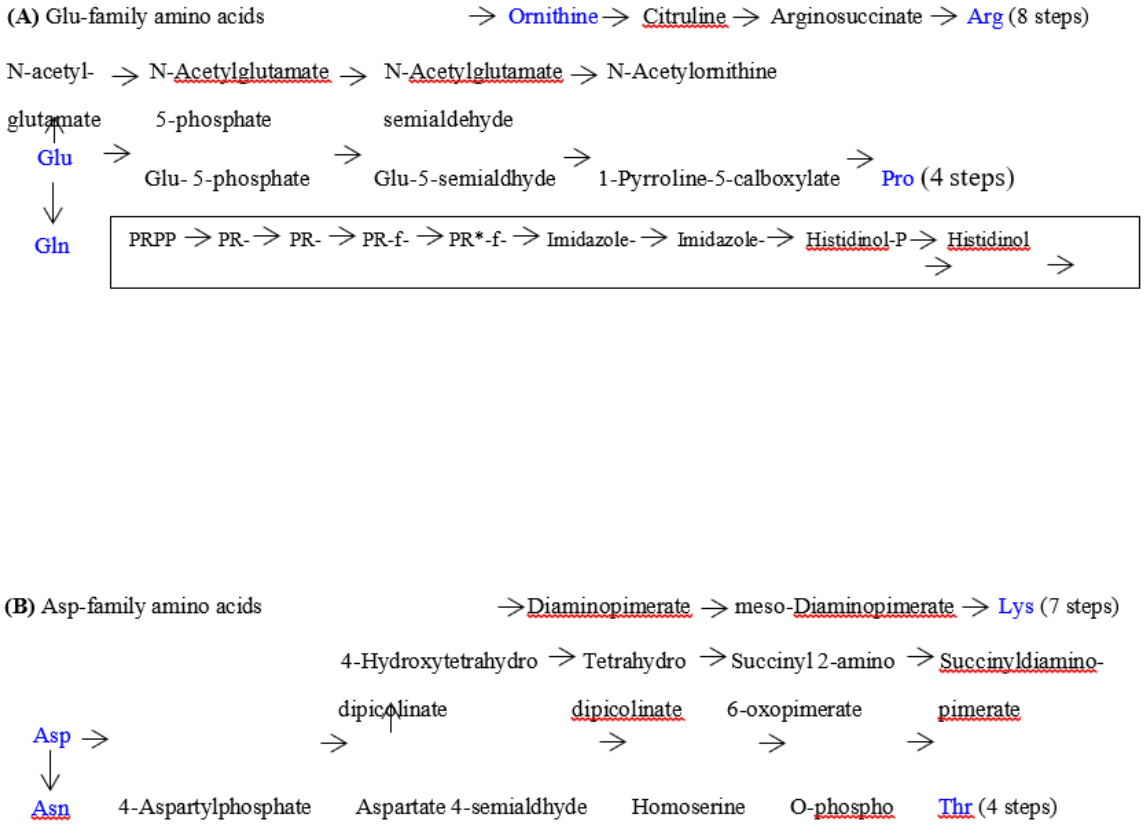

Figure 5. (A) Synthetic pathways of three amino acids (Pro, Gln, Arg) encoded by C-start codons. The three amino acids are synthesized from a precursor amino acid, Glu, encoded by GAR codons. His synthetic pathway is shown in an inlet. **(B)** Amino acid synthetic pathways of three amino acids (Thr, Gln and Lys) encoded by A-

 doi:10.20944/preprints202502.1594.v1

start codons. The three amino acids are synthesized from a precursor amino acid, Asp, encoded by GAY codons. Note that Arg and Lys synthetic pathways are shown in two lines.

Three amino acids, Thr, Gln and Lys, out of five amino acids encoded by A-start codons are synthesized through metabolic pathways using Asp as a precursor molecule (Figure 5) [6]. On the other hand, metabolic pathway for Ile synthesis is composed of 2-aceto-2-hydroxybutanoate synthetic pathway from pyruvate and the four-steps reaction pathway commonly using enzymes for Val synthesis (Figure 2) [6]. The rest of amino acid, Met, is synthesized at five steps from Ser via Cys (Figure 6) [6]. From the considerations, it can be concluded from a viewpoint of the coevolution theory that both C- and A-start codons were incorporated into a genetic code, after GNS-code was established, because two amino acids, Asp and Glu, are encoded by two G-start codons, GAC and GAG, respectively.

On the contrary, three aromatic amino acids, Phe, Tyr and Trp, out of five amino acids encoded by U-start codons, are produced through a branched pathway starting from phosphoenol pyruvate and erythrose 4-phosphate, which are synthesized through glycolysis and pentose-phosphate cycle, respectively (Figure 6) [6]. Ser is synthesized from 3-phophoglycerate, which is produced through metabolic pathway starting from glycelaldehyde on glycolysis. Cys is produced two reaction steps from Ser (Figure 6) [6]. Thus, the four amino acids except Cys are synthesized without using a precursor amino acid. Therefore, it is impossible to determine the time, when the amino acids encoded by U-start codons were introduced into what code previously existed, based on only the coevolution theory. In other words, the amino acids except Cys encoded by U-start codons are synthesized independently of any other amino acid. Nevertheless, it could be assumed that the amino acids encoded by U-start codons must be captured at the last stage after G-, C- and A-start codons were formed. The reason is because three aromatic amino acids are synthesized by reactions using more than ten steps, of which formation is expected to be quite difficult (Figure 6) [6].

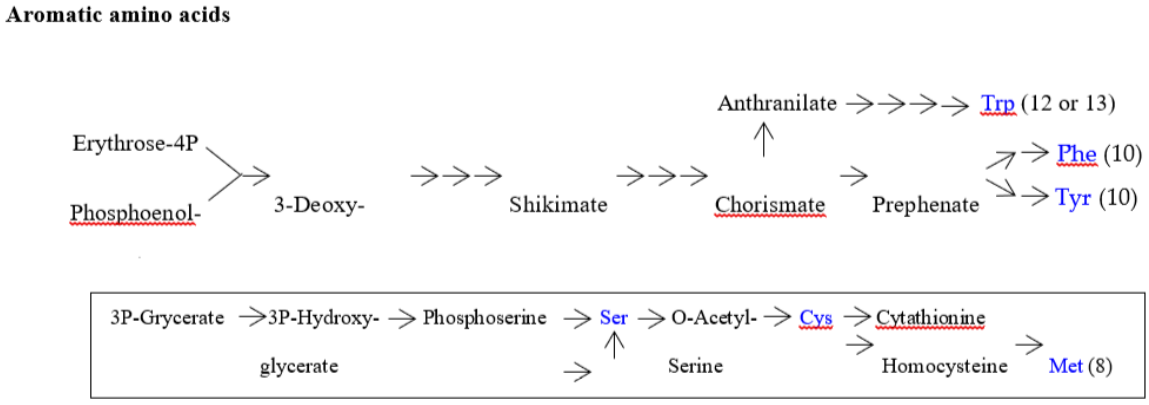

Figure 6. Synthetic pathways for three aromatic amino acids, Trp, Phe and Tyr. The amino acids are synthesized by complex metabolic reactions composed of more than 10 steps. It can be understand that Trp synthetic pathway was formed by branching out from chorismate and, therefore, Trp synthetic pathway was formed after completion of Phe/Tyr synthetic pathway. Boxed metabolic pathways show Met synthetic pathway from Ser via Cys. It is supposed that Ser synthetic pathway from Gly was added at a later time to replenish the shortage of Ser.

6.2. Characteristics and Differences of RNY code and SNS code

In this Subsection, the characteristics of the two genetic code, SNS code and RNY code, are summarized, because, by doing so, it would become possible to determine which code is appropriate as an intermediate code connecting the first GNC genetic code with universal genetic code.

1. Both SNS code and RNY code are consistent with the idea that the genetic code originated from the GNC primeval genetic code, because GNC code is contained in the two codes (Figure 4B,C).

 doi:10.20944/preprints202502.1594.v1

2. SNS code is symmetrical code between codon on sense strand and anticodon on antisense strand and is composed of ten amino acids and sixteen codons. Two acidic amino acids, Asp and Glu, and two basic amino acids, His and Arg, are contained in the SNS code. Two basic amino acids, His and Arg, and one highly hydrophobic amino acid, Leu, which are relatively complex amino acids, are contained in the SNS code. (Figure 4B).

3. RNY code is also symmetrical code between codon and anticodon and is composed of eight relatively simple amino acids except Ile and sixteen codons. No basic amino acid and only one acidic amino acid, Asp, is contained in the RNY code (Figure 4C).

4. In the case of SNS code, it is considered that the order of codon capture was GNC--GNG--CNC--CNG codons. On the contrary, it is supposed that codons were captured in order of codons GNC--GNU--ANC--ANU to form RNY code.

The reason, why such evolutionary process of the genetic code can be assumed as advanced in the sequence of codon captures, is because it can be supposed that in the case of SNS code, CNC codons complementary with GNG codons were captured into GNS (GNC + GNG) code upon entirely new gene formation from antisense strand [10,18]. On the other hand, in the case of RNY code, ANC codons complementary to GNU codons were introduced into GNY (GNC + GNU) code (Figure 7). Therefore, both cases are consistent with the idea that entirely new genes would be generated from antisense strand of GC-rich gene.

Figure 7. The order of a new codon usage. **(A)** In the case of SNS code, it is supposed that formation of SNS code started from GNC code and was established by creation and addition of new codons to the GNC code in order of GNG codon formation on sense strand and CUC codon formation on antisense strand. **(B)** In the case of RNY code, it is assumed that RNY code started from GNC code and was established by creation and addition of new codons to the GNC code in order of GNU codon formation on sense strand and AUC codon formation on antisense strand.

6.3. Evidences showing that SNS code but not RNY code was used as an intermediate code

From here, discuss which code, SNS code or RNY code, was used as an intermediate code connecting the first GNC code with the universal genetic code. Then, two evidences showing that SNS code was used earlier than RNY code are given.

1. As can be seen in Figure 8, codon sequences on sense strand and antisense strand of modern GC-rich genes are similar to $(SNS)_n$ [18]. The fact is one evidence showing that $(SNS)_n$ were certainly used as a genetic sequence in the ancient days, because it is considered that the remnants of SNS codon usage remain in modern GC-rich genes and in its antisense codon sequences still now. Contrary to that, RNY pattern cannot be found out in any region of modern microbial genes (Figure 8).

2. Furthermore, the remnants of evolution of tRNA were found out by analyses of anticodon-stem sequences of *Pseudomonas aeruginosa* tRNAs, which actually realize the correspondence relationships between a codon and an amino acid [14].

The evidences above clearly indicate that the genetic code, which originated from GNC code, evolved to universal genetic code via SNS code.

In addition, there is another defect in evolutionary process passing through RNY code, because GC content of genes changes largely from average 83% GC content of GNC code to average 50% GC content of RNY code. On the contrary, GC content of genes slowly changes from GNC (average 83% GC content)-SNS code (average 83% GC content) and to universal genetic code (average 50% GC

 doi:10.20944/preprints202502.1594.v1

content) taking a long time. In addition, the field for entirely new gene formation can be remained in antisence codon sequence of GC rich genes [18]. However, it must be stated that the field for entirely new gene formation should be lost upon the change from GNC code to RNY code.

Thus, it can be concluded that not RNY code but SNS code was used as an intermediate code from the evidences showing that there was an evolutionary route from GNC code to SNS code.

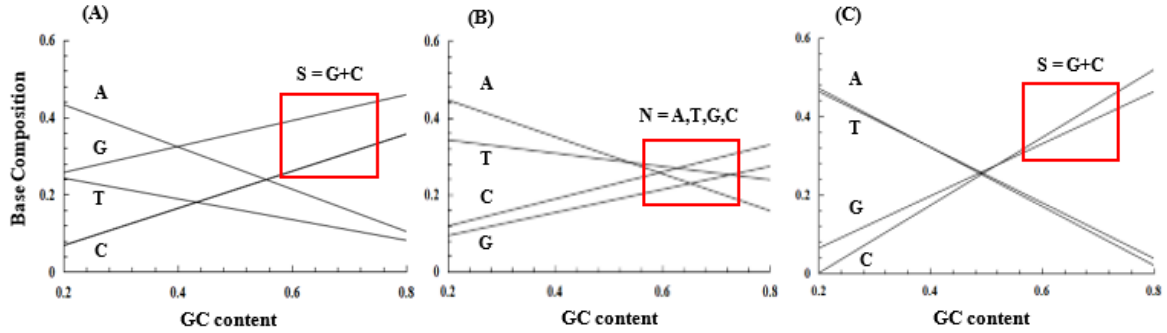

Figure 8. The evidence that remnant of SNS code remains in GC-rich region of modern microbial genomes. Note that remnant of GNC code also remains in the GC-rich region. On the contrary, such remnant of RNY code cannot be found any region of the microbial genomes.

6.4. Decoding SNS Code and RNY Code by tRNA

All the sixteen codons in SNS code can be decoded by tRNA without wobbling recognition. On the contrary, it would be necessary to decode RNY code with eight pairs of RNU and RNC codons as using wobble recognition at the third codon position. However, a complex mechanism, such as ribosome, should be required to enable the wobble recognition between two bases (U and C) at the third codon position. Therefore, it would be quite difficult to decode RNY code without elaborate mechanism in the ancient days, because base pairs would be generally formed as a result of seeking for the most stable position (Watson-Crick base pair formation). Therefore, it is considered that it would be actually impossible to use RNY code as an intermediate code. In addition, it must be indicated that the number of amino acids, which can be used in proteins, does not increase even by using the wobble recognition.

7. Limits of the Coevolution Theory

7.1. The Reason Why ANN Code Was Introduced into SNN Code After SNN Code Was Formed

Three amino acids, Pro, Gln and Arg, which are synthesized from Glu, are contained in SNS code. Contrary to that, Thr, Asn, and Lys, are synthesized through metabolic pathways starting from a more simple acidic amino acid, Asp, than Glu. Therefore, it would be difficult to explain the reasons, why A-start codons were captured after SNN code was established. Stating this more concretely, it should be naturally more difficult to form the SNS code using three more complex amino acids than the three amino acids encoded by A-start codons. If, nevertheless, the genetic code evolved in order of GNC code, SNS code, SNN code and (SNN + ANN) code as shown in Figure 9, it must be stated that there is a contradiction. However, the contradiction could be solved by understanding the facts described below.

GNC--GNS(C+G)--(GNS+CNC)---**SNS**(GNS+CNS)-------(SNN + ANN)----------**U.G.C.**(V(G+C+A)NN+UNN
[GADV] [GADV + E] [GADVE + LPHA] [GADVE-LPHA + Q] [GADVE-LPHAQ + I(M)TNK] [GADVE-LPHAQ I(M)TNK + FSYCW]

Figure 9. Evolutionary process deduced from the coevolution theory and the GNC-SNS primitive genetic code hypothesis. See *Subsections, 6.2 and 6.3* about the reasons why the evolutionary pathway pass through SNS code. Capital letters in square brackets show amino acids written with one letter symbol. Alphabets written in blue letters indicate amino acids, which were newly captured into the previously existed code.

 doi:10.20944/preprints202502.1594.v1

1. A large amount of Asp was required for syntheses of purine and pyrimidine nucleotides (Figure 3) [6], which are monomers of RNA as a genetic information carrier. Therefore, it is supposed that Asp did not accumulate at a large amount in cell structure. On the contrary, Glu was not required as molecules for syntheses of other important molecules. Therefore, Glu easily accumulated at a large amount in cell structure. The contradiction could be resolved by understanding that Glu, which accumulated at a large amount in the cell earlier than Asp. Thus, SNS code was established earlier than formation of RNN code.

2. Furthermore, it must be answered to the question, why the three amino acids could be synthesized at a large amount from Asp and could be used in ANN code, although it was later than establishment of the SNS code. The answer is because a new route for Asp synthesis starting from pyruvate was invented (Figure 3) [6] and Asp accumulated at a sufficiently large amount to produce new amino acids used in ANN code. Consequently, Asn, Thr and Lys could be used in a new genetic code as (SNN + ANN) code (Figure 9).

7.2. The reason why it could not be determined which one of two evolutionary pathways passed through based on only coevolution theory or amino acid metabolism

Here, consider the reason why it could not be determined which one of two evolutionary pathways passed through, one is the incorporation route of five amino acids (Leu, Pro, His, Gln and Arg) including three amino acids (Pro, Gln, Arg) encoded by CNS code into GNS code encoding [GADV+E]-amino acids, and the other is the incorporation route of four amino acids encoded by ANY code (Ile, Thr, Asn, Ser) including two amino acids (Thr, Asn) into GNY code encoding [GADV]-amino acids. The reason is because the coevolution theory can only suppose that a new metabolic pathway could be formed using a metabolite, which was synthesized through previously existed metabolic pathway and accumulated at a sufficiently large amount in cells. Therefore, the theory can only deduce that a new metabolic pathway should be formed later than the pathway, through which a precursor molecule was supplied. Therefore, it would be important to understand the order of metabolic pathway formation is not always determined by the coevolution theory, although the coevolution theory is valid and is undoubtedly one of quite important theories [1–5].

It is considered that the synthetic pathways between two amino acids, for example Ser-Gly (Figure 6) and Gly-Thr, were invented to refill one amino acid shortage with another amino acid accumulating at a large amount and vice versa. Thus, it must be noted that there are some cases, in which new metabolic pathways synthesizing the same amino acid are formed in addition to the previously formed pathway.

7.3. Examples appropriate to the coevolution theory

Of course, there are many typical examples for the coevolution theory. The following amino acid pairs, which are frequently given as examples of precursor amino acid-product amino acid, are discussed. The two cases, Asp-Asn, Glu-Gln, are typical examples supporting the coevolution theory (Figure 2). Both Asp and Glu are precursor amino acids and both Asn and Gln are product amino acids, respectively. In addition, Asn and Gln are synthesized from Asp and Glu at one step reaction.

In the case of Asp-Glu, the precursor molecule of Asp synthesis, oxaloacetate, is used for Glu synthesis (Figure 2). In another example, Ala-Val, the precursor molecule of Ala synthesis, pyruvate, is used for Val synthesis (Figure 2). Therefore, it can be considered that the two pairs, Asp-Glu and Ala-Val, are also examples suitable to the coevolution theory. Furthermore, the idea of coevolution theory can be applied to all metabolites including amino acids. Thus, it can be stated that the coevolution theory is one of brilliant ideas for studies on the origins and evolution of the genetic code and metabolic pathways, although there are some limits in the coevolution theory described above.

8. Discussion

Significance of the coevolution theory is first described and it was confirmed that the theory is valid as an example of formation process of GNC primeval genetic code. After the conditions, which

 doi:10.20944/preprints202502.1594.v1

must be satisfied when the origin and evolutionary process of the genetic code are analyzed by using the coevolution theory, were confirmed, the universal genetic code, which originated from GNC code, evolved through what type of an intermediate code were analyzed from the viewpoint of the coevolution theory and/or amino acid synthetic pathways. In the analyses, it could not be determined which code, SNS code or RNY code, was used as an intermediate code to the universal genetic code. Therefore, it was determined that the universal genetic code originated from GNC primeval genetic code and evolved through SNS code based on characteristics and their differences of SNS code and RNY code. Thereafter, the reasons, why SNS code was formed earlier than RNY code, are discussed and it has been concluded that there was not the period, when RNY code itself was used. Furthermore, the reason, why it could not be determined which code, SNS code or RNY code, was used as an intermediate code, is also discussed.

In the analyses, it was found that many amino acids can be synthesized using an acidic amino acid, Glu or Asp, as a precursor molecule. The reason, why acidic amino acids can be used for synthesis of other amino acids, is explained as that structures of acidic amino acids, Asp and Glu, are comparatively simple and it was convenient to use the amino acids, which accumulated at a high amount in cells at early stage of the emergence of life.

Further, the reason, why structures of acidic amino acids are simpler than those of basic amino acids, is discussed. The reason would be because it was easy to synthesize simply structured acidic amino acids with carboxylic residue and also because positive charges, which are not contained in acidic amino acids, could be substituted by divalent metal ions, as Mg^{2+}, Mn^{2+}, Ca^{2+}, Cu^{2+} and etc. Contrary to that, the reason, why basic amino acids, which are used in the genetic code, have relatively complex structure, would be because basic amino acids were synthesized much later than acidic amino acids and it was necessary to avoid to compete with acidic amino acids having a simple structure, except charge of side chain.

Lastly, discuss the time when Met was synthesized and incorporated into a previously existed genetic code. The reason is because Met encoded by A-start codon (AUG) is synthesized by using Cys encoded by U-strat codons (UGY) as a precursor amino acid (Figure 6). This means that the codon for Met should be incorporated into a genetic code after Cys synthetic pathway was formed according to the coevolution theory. However, on the other hand, Met encoded by A-start codon should be incorporated into a genetic code before Cys capture into a genetic code according to the conclusion that A-start codons should be captured earlier than U-start codons as described in this article (Figure 9). This means that there is another contradiction between the times when two amino acids, Met and Cys, were incorporated into a genetic code. The reason can be explained as follows. Met was synthesized at the last stage of evolutionary process of the genetic code or after synthetic pathway of Cys encoded by U-start codon was completed. Further, it can be reasonably explained as that, at that time, when Met synthetic pathway was formed, the used amount of AUG codon assigned into Ile was small, and therefore, the change of codon assignment from Ile to Met could be carried out without any large obstacle. The validity could be supported by the fact that the usage amount of AUG codon for Met is quite small even at this time point.

Needless to say, the origins and evolution of the genetic code relate not only the metabolic pathways for amino acid synthesis but also tRNA and aminoacyl tRNA synthetase [15,19–24]. Therefore, it is expected that studies on the origins and evolution of the genetic code make great progress through the comprehensive studies about some members related to the genetic code.

References

1. Wong, J.T. A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci. USA*. **1975**, 72. 1909-1912.
2. Wong, J.T.; Ng, S.K.; Mat, W.K.; Hu, T.; Xue, H. Coevolution Theory of the Genetic Code at Age Forty: Pathway to Translation and Synthetic Life. *Life (Basel)* **2016**, 6(1), 12.
3. Di Giulio, M. The coevolution theory of the origin of the genetic code. *J Mol Evol.* **1999,** 48. 253-255.
4. Di Giulio, M. An extension of the coevolution theory of the origin of the genetic code. *Biol Direct*. **2008**, 3:37

5. Di Giulio, M. Theories of the origin of the genetic code: Strong corroboration for the coevolution theory. *Biosystems*. **2024** 239, 105217.
6. KEGG Pathway Dtabase. https://www.genome.jp/kegg/pathway.html
7. Frank, A.; Froese, T. The Standard Genetic Code can Evolve from a Two-Letter GC Code Without Information Loss or Costly Reassignments. *Orig. Life Evol. Biosph.* **2018**, 48, 259–272. https://doi.org/10.1007/s11084-018-9559-4
8. Higgs P.G. A four-column theory for the origin of the genetic code: Tracing the evolutionary pathways that gave rise to an optimized code. *Biol Direct*, **2009**. 24, 4–16
9. Ikehara, K.; Omori, Y.; Arai, R.; Hirose, A. A novel theory on the origin of the genetic code: a GNC-SNS hypothesis. *J. Mol. Evol*. **2002**, 54, 530–538.
10. Ikehara, K. *Towards Revealing the Origin of life.—Presenting the GADV Hypothesis*; Springer Nature, Gewerbestrasse: Cham, Switzerland, **2021**.
11. Ikehara, K. Protein ordered sequences are formed by random joining of amino acids in protein 0^{th}-order structure, followed by evolutionary process. *Orig. Life Evol. Biosph*. **2014**, 44, 279–281.
12. Ikehara, K. Why were [GADV]-amino acids and GNC codons selected and how was GNC primeval genetic code established? *Genes* **2023**, 14, 375.
13. Taghavi, A.; van der Schoot, P.; Berryman, J.T. DNA partitions into triplets under tension in the presence of organic cations, with sequence evolutionary age predicting the stability of the triplet phase. *Q. Rev. Biophys.* **2017**, e15.
14. Ikehara, K. The origin of tRNA deduced from *Pseudomonas aeruginosa 5′* anticodon-stem sequence: Anticodon stem-loop hypothesis. *Orig. Life Evol. Biosph*. **2019**, 49, 61-75.
15. Lei L.; Burton, Z.F. Evolution of Life on Earth: tRNA, Aminoacyl-tRNA Synthetases and the Genetic Code. *Life (Basel)*, **2020**, 10(3), 21.doi: 10.3390/life10030021.
16. Watson, J.D.; Hopkins, N.H.; Roberts, J.W.; Steitz, J.A.; Weiner, A.M. *Molecular Biology of the Gene. 4th ed.* Vol. 1, The Benjamin/Cummings Publishing Company, Inc. Menlo Park, California, U.S.A. **1987**.
17. Zamudio, G.S.; José, M.V. On the Uniqueness of the Standard Genetic Code. *Life (Basel).* **2017**, *7*(1), 7
18. Ikehara, K.; Amada, F.; Yoshida, S.; Mikata, Y.; Tanaka, A. A possible origin of newly-born bacterial genes: Significance of GC-rich nonstop frame on antisense strand. *Nucl. Acids Res.* **1996**, 24:4249–4255.
19. Lei, L.; Burton, Z.F. Evolution of the genetic code. *Transcription*, **2021**, 12(1), 28-53.
20. Kim, Y.; Opron, K. Burton, Z.F. A tRNA- and Anticodon-Centric View of the Evolution of Aminoacyl-tRNA Synthetases, tRNAomes, and the Genetic Code. *Life (Basel)*, **2019**, 9(2), 37.
21. Rogers, S.O. Evolution of the genetic code based on conservative changes of codons, amino acids, and aminoacyl tRNA synthetases, *J. Theor. Biol.* **2019**, 466, 1-10.
22. Foltan, J.S. tRNA genes and the genetic code. J. Theor. Biol. **2008,** 253(3), 469-482.
23. Jackman. J.E.; Alfonzo, J.D. Transfer RNA modifications: nature's combinatorial chemistry playground. *Wiley Interdiscip. Rev. RNA*. **2013** 4(1), 35-48.
24. Yared, M.J.; Marcelot, A.; Barraud, P. Beyond the Anticodon: tRNA Core Modifications and Their Impact on Structure, Translation and Stress Adaptation. *Genes (Basel)*. **2024**, 15(3), 374.