

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

186,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Origin of the Genetic Code and Genetic Disorder

Kenji Ikehara

*The Open University of Japan, Nara Study Center
International Institute for Advanced Studies of Japan
Japan*

1. Introduction

Genetic disorders are illnesses caused by abnormalities in genetic sequences and the chromosome structures. Most base substitutions, which may lead to genetic disorders, would be repressed to a low level as affecting only one person in every thousands or millions by replication repair systems and by robustness of the genetic code, which is discussed in this Chapter. But, once persons were suffered by the genetic disorders, they would probably get serious diseases during their lives. In addition, it is quite difficult to recover the substituted bases causing the genetic diseases to original bases, after persons were suffered by the rarely occurring genetic disorders. This makes a quite big problem of the genetic disorders from a stand point of medical treatment.

The mutations causing the genetic disorders are scattered throughout genes and their neighboring regions as shown in Figure 1 (A). It is also known that many genetic diseases are induced by single-base substitutions or missense mutations including nonsense mutations in genetic regions encoding amino acid sequences of proteins. For instance, sickle-cell anemia, one of the classical genetic disorders, is caused by a one-base replacement at the sixth codon of the hemoglobin β -globin gene, from A to U, which results in one amino acid substitution from glutamic acid to valine, producing an abnormal type of hemoglobin called hemoglobin S (Figure 1 (B)). Hemoglobin S distorts the shape of red blood cells due to hemoglobin aggregation in the cells, especially when exposed to low oxygen levels, resulting in anemia giving a patient malaria resistance. Phenylketonuria (PKU), adenosine deaminase (ADA) deficiency and galactosemia are also caused by one-base replacements in genes of phenylalanine hydroxylase, adenosine deaminase and galactosidase, respectively (Table 1). Of course, deletion and insertion of a small number of bases causing frameshift mutations in a genetic sequence encoding protein may also affect normal life activities, because the frameshift mutation induce a change to different amino acid sequences following the mutation site. Base substitutions also may occur in transcriptional and translational control regions, splicing sites and so on, which affect various functions for gene expression leading to synthesis of lower or higher amounts of proteins than normal level, resulting in many kinds of genetic diseases (Figure 1 (A)).

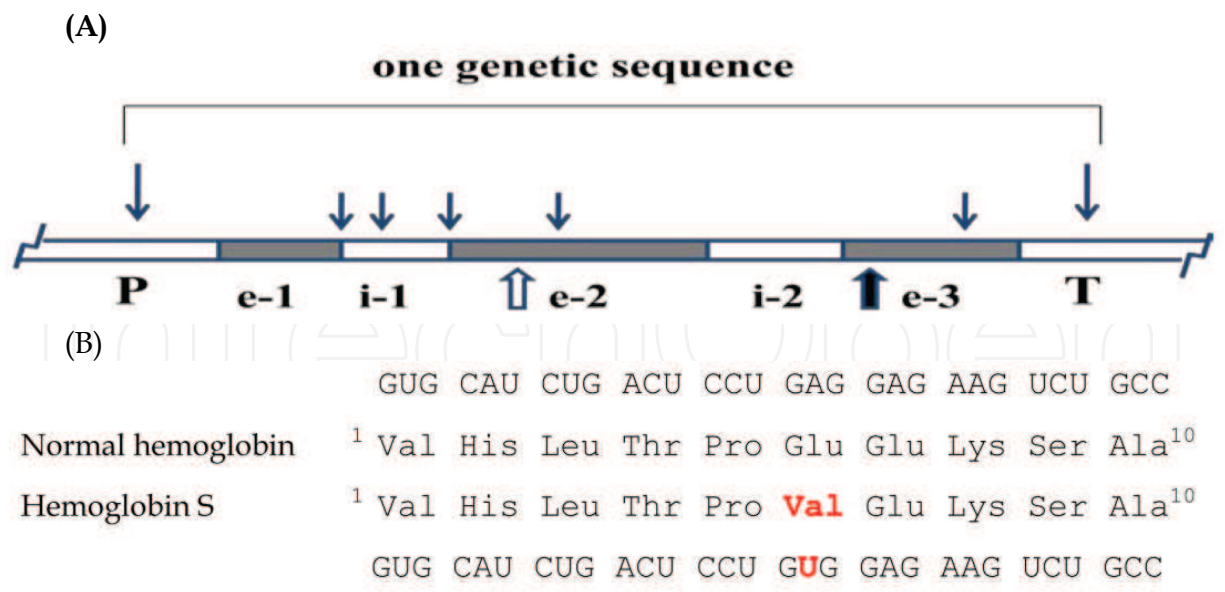


Fig. 1. (A) Possible mutation sites, which may affect various functions for gene expression and catalytic functions of proteins. Dark and white horizontal bars indicate exons encoding amino acid sequences of a protein and introns without genetic information for protein synthesis, respectively. Capital letters, P and T, mean a promoter for transcription initiation and a terminator required for termination of mRNA synthesis, respectively. Thick upward open and closed arrows and thin downward arrows indicate insertion and deletion of DNA sequences, and one-base substitutions, respectively. (B) Amino acid replacement observed in a classical and well-known genetic disorder, sickle cell anemia. Red letters indicate replacements of amino acid and base of the genetic mRNA sequence

Genetic Disorder	Inheritance	Gene
Hailey-Hailey Disease	Autosomal dominant	ATP2C1
Adenosine deaminase deficiency	Autosomal recessive	ADA
Thalassemia		globins
Alstrom Syndrome		ALMS1
Tangier Disease		ABCA1
Phenylketourea		PAH
Galactosemia		GALT
Aicardi-Goutieres syndrome	X-link dominant	RNAses
Bernard-Soulier syndrome		GPIs
Wiskott-Aldrich syndrome	X-link recessive	WASp
Fabry Disease		α-Gal A
Ornithine transcarbamoylase deficiency		OTC

Table 1. Examples of representative genetic disorders caused by one-base replacements on genetic sequences encoding amino acid sequences of proteins

Base substitutions might occur on every gene encoding functional proteins on a whole genome. In fact, about ten thousands genetic diseases are already known until now, out of which several genetic disorders caused by one-base replacements or monogenic disorders are described in Table 1.

In this Chapter, I will discuss on genetic disorders, which are caused by one-base replacements in coding regions, because I would like to discuss on relationships among robustness of the universal genetic code, base substitutions in codons and genetic disorders from a stand point of the origin of the genetic code. Term of “the universal genetic code”, which is widely used in extant organisms, is used in this Chapter, instead of “the standard genetic code”, which is used in many textbooks of in the fields of biochemistry and molecular biology since discoveries of non-universal genetic codes in mitochondria of mammals, protozoa and some bacteria. That is because I would like to emphasize that almost all organisms on this planet have actually used the genetic code. I believe that understanding on the relationship between the robustness and base substitutions will contribute to discovery of proper methods for treatments of many genetic disorders in a future.

Amino acid substitutions not largely affecting normal protein function are observed, as it is known as single nucleotide polymorohisms in the case of human beings. But, amino acid substitutions of mammals evolving at a quite slow rate due to a long generation time, such as about 25 years in the case of human, have occurred at a comparatively low frequency. On the other hand, amino acids of microbial proteins have been substituted at a high frequency without largely affecting protein functions. That is because evolution rate of microbial proteins is quite large due to the enormously large cell number and a quite short division time, such as about 20-30 minutes in the case of *Escherichia coli*. Therefore, it would be suitable to compare an amino acid sequence of a microbial protein with the homologous amino acid sequence in order to investigate amino acid substitutions occurring without largely affecting the protein function in a wide range as shown in Figure 2.



Fig. 2. Alignment of two amino acid sequences of small homologous single-stranded DNA binding proteins, from *Aquifex aeolicus* (147 amino acids) and *Carboxydotherrmus hydrogenoformans* (142 amino acids). Red bold and black letters indicate substituted and conserved amino acids between the two amino acid sequences, respectively. Hyphen (-) means amino acid position deleted from one amino acid sequence. Homology percent between the two single-stranded DNA binding proteins, which were obtained from GeneBank at <http://www.ncbi.nlm.nih.gov/genbank/>, is 38%

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A		0,0	4,0	6,0	0,0	1,2	2,0	2,0	1,0	2,0	2,0	4,0	1,0	2,0	3,1	6,0	2,0	4,1	0,0	3,0
C	0,0		0,0	0,0	0,0	0,0	0,0	1,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
D	0,0	1,0		5,1	1,0	1,0	0,0	0,0	4,0	1,0	2,0	2,0	0,0	3,0	0,0	2,0	2,1	0,0	0,0	0,0
E	1,0	0,0	1,5		1,1	0,1	0,0	1,1	5,0	0,1	1,0	1,1	1,1	3,0	3,2	2,3	2,1	1,0	0,0	2,0
F	0,0	0,0	0,0	0,0		0,0	0,0	2,3	0,0	1,1	0,0	0,0	0,0	1,0	1,1	0,0	0,0	1,0	0,0	5,0
G	1,0	0,0	1,0	1,0	0,0		0,0	0,0	5,0	0,0	0,0	3,1	0,0	2,1	1,1	2,0	1,0	0,0	0,0	1,0
H	1,0	0,0	1,1	1,0	0,0	1,0		0,0	0,0	0,0	2,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,0
I	0,0	0,0	0,0	1,0	0,0	0,0	0,0		0,0	3,3	1,0	0,0	0,1	0,0	0,0	0,0	0,0	7,3	0,0	1,0
K	2,0	0,0	2,1	4,0	1,0	0,0	1,0	1,1		0,0	0,0	0,0	2,0	0,1	3,0	0,1	0,1	1,2	0,0	1,0
L	1,0	0,0	0,0	0,0	3,3	1,0	0,0	14,0	0,0		5,1	0,0	0,0	2,0	1,0	0,0	1,2	5,1	0,0	2,0
M	0,0	0,0	0,0	0,0	0,0	0,0	0,0	3,0	0,0	5,1		0,0	0,0	1,0	0,0	0,0	0,0	2,0	0,0	1,0
N	0,0	0,0	2,2	1,1	0,0	2,0	0,0	0,0	1,0	0,0	0,0		0,0	1,0	0,0	0,0	1,1	0,0	0,0	0,0
P	1,1	0,0	1,0	1,0	0,0	2,0	0,0	1,0	1,0	1,0	0,0	2,0		0,0	2,0	2,0	1,0	1,0	0,0	1,0
Q	0,0	0,0	1,0	5,0	0,0	0,0	2,0	0,0	2,1	0,0	0,0	1,0	0,1		3,0	0,0	2,1	0,0	0,0	0,0
R	0,0	0,0	3,0	4,1	0,0	1,0	0,0	2,0	17,1	1,0	0,0	6,0	1,1	2,0		3,0	1,0	1,0	1,0	0,0
S	3,0	1,0	4,0	0,0	0,0	0,0	1,0	1,0	5,0	1,0	0,0	5,0	0,0	1,2	1,1		3,2	2,0	0,0	1,0
T	2,0	0,0	1,0	0,0	0,0	1,0	0,0	3,0	0,0	2,0	2,0	5,0	0,0	0,0	0,1	6,0		3,1	0,0	0,0
V	4,1	0,0	0,0	2,1	1,1	2,0	1,0	15,0	1,0	5,0	2,0	1,0	1,0	1,0	0,0	0,0	4,0		0,0	0,1
W	2,1	0,0	0,0	0,0	1,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,0	0,0	0,0	0,0	0,0	0,0		0,1
Y	1,0	0,0	1,0	0,0	3,1	1,0	1,1	1,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,1	0,1	

Protein	1st	2nd	3rd	1,2	1,3	others
RelA	119	93	13	10	8	154
SS-DNA.B	21	13	6	2	5	29

Fig. 3. The numbers of permissible amino acid substitutions observed between two pairs of homologous proteins, from *S. coelicolor* (left column) and to *S. aureus* (top row) RelA proteins (the numbers at the left side) and from *A. aeolicus* (left column) and to *C. hydrogenoformis* (top row) single-stranded DNA binding proteins (the numbers at the right side). Amino acid replacements upon base substitutions at the first, the second and the third codon positions are written in blue, yellow and red color boxes, respectively. Green, orange and white boxes indicate amino acid replacements induced by base substitutions at the first or the second codon positions, at the first or the third codon positions and other base substitutions, respectively. The base substitutions at the respective codon positions were deduced from amino acid replacements between two homologous proteins, which were occurred by one-base substitutions. The amino acid sequences, which were used for alignment, were obtained from GeneBank at <http://www.ncbi.nlm.nih.gov/genbank/>

As seen in Figure 2, many amino acid substitutions are observed between two homologous single-stranded DNA binding proteins. The amino acid substitutions caused by base substitutions at the first codon position were observed more than those caused by base substitutions at the second codon position (see the Table given in Figure 3). Similar results were obtained from amino acid substitutions between two large homologous stringent response proteins, *Streptomyces coelicolor* RelA and *Staphylococcus aureus* RelA (Figure 3). It can be interpreted as that amino acids with similar chemical and physical properties are arranged in the same column in the genetic code table at a comparably high probability (Table 2 (A), (B), (C) and (D)).

The universal genetic code is redundant and has a highly non-random structure. Typically, when nucleotide at the third codon position differs from the corresponding one, both codons encode the same amino acids at a high probability, due to the degeneracy of the genetic code at the third codon position. In addition, codons, of which nucleotide at the first codon position differs from each other, usually encode amino acids with different but rather similar chemical/physical properties.

(A)						(B)					
Hydropathy						α -Helix					
	U	C	A	G			U	C	A	G	
U	Phe	Ser	Tyr	Cys	U	U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C		Phe	Ser	Tyr	Cys	C
	Leu	Ser	Term	Term	A		Leu	Ser	Term	Term	A
	Leu	Ser	Term	Trp	G		Leu	Ser	Term	Trp	G
C	Leu	Pro	His	Arg	U	C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C		Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A		Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G		Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U	A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C		Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A		Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G		Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U	G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C		Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A		Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G		Val	Ala	Glu	Gly	G

Table 2. Color representation of chemical/physical properties, of amino acids based on the values described in Stryer’s “Biochemistry” (Berg *et al*, 2002). (A) hydrophobicities and (B) α -helix propensities of amino acids in the universal genetic code table. Letters in red, yellow and blue boxes represent amino acids with large, middle and small hydrophobicities, and the corresponding degrees of α -helix propensities, respectively

It can be seen in Table 2 that amino acids encoded by 16 codons in the same column are located in the same or two colored boxes at a high probability, such as two columns from left side of Table 2 (A) and one column at the most left side of Table 2 (D). Contrary to that,

no row with the same color boxes is observed in Table 2 (A), (B), (C) and (D). This means that amino acids with similar chemical/physical properties are arranged in the same column, but those with rather different chemical/physical properties are arranged in the same rows at high probabilities. As a result, it makes the genetic code to be highly robust to the change of protein functions upon base substitutions in protein coding sequences, especially at the third and the first codon positions of genetic sequences. My original GNC-SNS primitive genetic code hypothesis on the origin and evolution of the genetic code (Ikehara, et al., 2002), which will be described in Section 3, can explain reasonably the robustness of the genetic code, which might stem from the origin and evolutionary processes. N and S mean either of four bases (A, U/T, G and C) and G or C, respectively.

(C)						(D)					
β-Sheet						Turn/Coil					
	U	C	A	G			U	C	A	G	
U	Phe	Ser	Tyr	Cys	U	U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C		Phe	Ser	Tyr	Cys	C
	Leu	Ser	Term	Term	A		Leu	Ser	Term	Term	A
	Leu	Ser	Term	Trp	G		Leu	Ser	Term	Trp	G
C	Leu	Pro	His	Arg	U	C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C		Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A		Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G		Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U	A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C		Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A		Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G		Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U	G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C		Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A		Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G		Val	Ala	Glu	Gly	G

Table 2. (Contn'd). (C) β-sheet and (D) turn/coil structure propensities, of amino acids in the universal genetic code table. Letters in red, yellow and blue boxes represent large, middle, and small β-sheet and turn/coil propensities, respectively. Meanings of color boxes in Table (C) and (D) are the same as in Table (A) and (B), described above. Secondary structure (β-sheet; (C) and turn/coil; (D)) propensities of amino acids were obtained from Stryer’s “Biochemistry” (Berg *et al*, 2002)

2. Significance of the Genetic Code for life

The genetic code plays a quite important role in transfer of genetic information on DNA nucleotide sequence to amino acid sequence of a protein, such as enzyme and transporter of a chemical compound, *etc* (Figure 4). But, the genetic code has been generally regarded as a simple representation of the relationship between a genetic information or a codon composed of three bases (triplet) and an amino acid in a protein sequence as described in

representative text books, as Stryer's "Biochemistry" (Berg *et al*, 2002). It seems to me that the significance of the genetic code has been underestimated at the present time, judging from my original idea suggesting that protein 0th-order structures, which are specific amino acid compositions favorable for effectively producing water-soluble globular proteins even by random synthesis (see Section 4), are secretly described in the genetic code table (see Figure 7 in Section 3).

Genetic information, which is stored in base sequences or actually in codon sequences on DNA, is propagated from a parent to progeny cells through DNA replication. In parallel, the information is transformed into mRNA and successively into an amino acid sequence of a protein according to the genetic code, when necessary. Various organic molecules required to live are synthesized with enzyme proteins on metabolic pathways (Figure 4). Therefore, it is no exaggeration to say that the genetic code is much more significant for lives than genes and proteins, or that the genetic code is the most important facility in the fundamental life system. Understanding of the origin and evolutionary processes of the genetic code should be quite important to know a framework of the genetic code and a relationship between amino acid substitutions and one-base substitutions causing genetic disorders.

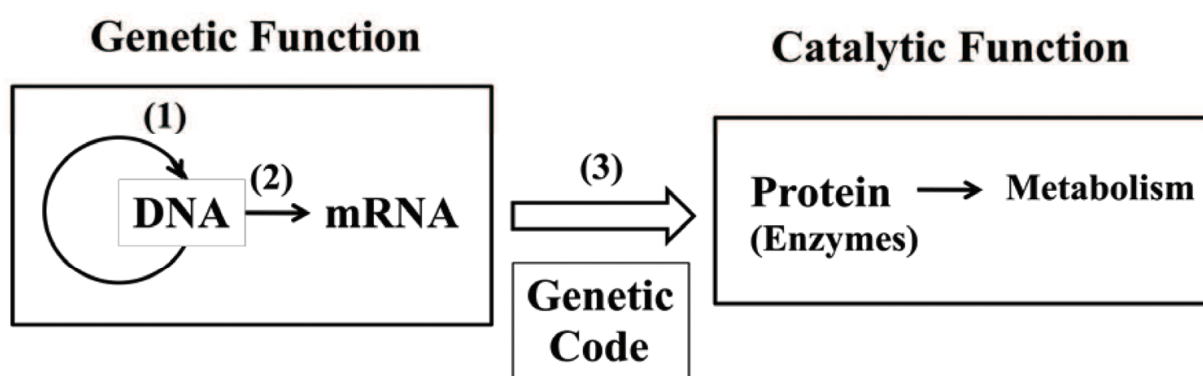


Fig. 4. Role of the genetic code playing in the fundamental life system of modern organisms, which is composed of genes, the genetic code and proteins (enzymes). Genetic code mediates between two main elements, genetic function composed of DNA (mRNA) and function carried out by proteinaceous catalysts (enzymes) forming chemical network or metabolism. Genetic information on DNA are transmitted to progeny cells by replication (Step 1), and transcribed into mRNA (Step 2) when necessary. Genetic information transferred into mRNA is translated to the corresponding amino acid sequence of a protein (Step 3) through genetic code mediating genetic information and catalytic function. The universal genetic code used by extant organisms on the earth is composed of 64 codons and 20 amino acids (see Table 2)

3. Origin of the Genetic Code (GNC-SNS primitive genetic code hypothesis)

Our studies on the origin of the genetic code were initiated from the search for a prospective spot on a DNA sequence, from which an entirely new gene encoding an entirely new functional protein will be created, when an extant organism using the universal genetic code has to adapt to a new environment. The spot was searched based on the six necessary conditions for producing water-soluble globular proteins as described below. The six conditions used for the search are hydropathy, α -helix, β -sheet and turn/coil formabilities,

acidic amino acid and basic amino acid contents of proteins, which were obtained as average values plus/minus standard deviations of water-soluble globular proteins in extant micro-organisms. From the results, it was found that non-stop frames, which appear on anti-sense strands of GC-rich genes (GC-NSF(a)s) at a high probability, have the strongest possibility to create entirely new genes, not new modified type of genes or homologous genes (Figure 5) (Ikehara et al., 1996). Where GC-NSF(a) means nonstop frame on antisense strand of GC-rich gene. That is because hypothetical proteins encoded by GC-NSF(a)s satisfied the six conditions and because the probability of non-stop frame (NSF) appearance on the GC-rich anticodon sequences was enough high (Ikehara, 2002).

The GC-NSF(a) hypothesis on creation of the first family genes under the universal genetic code led us propose subsequent theory on the origin of the genetic code as GNC-SNS primitive genetic code hypothesis (Ikehara et al., 2002). GNC and SNS represent four codons (GUC, GCC, GAC and GGC) and 16 codons (GUC, GCC, GAC, GGC, GUG, GCG, GAG, GGG, CUG, CCG, CAG, CGG, CUC, CCC, CAC and CGC), respectively. I describe the clues briefly below, from which the hypothesis was obtained. The first one is that base sequences of the GC-NSF(a)s were rather similar to the repeating sequences of SNS. The second one is that hypothetical proteins encoded by GNC code, a part of the SNS code, satisfied the four conditions (hydropathy, α -helix, β -sheet and turn/coil formabilities of proteins) for folding polypeptide chains into water-soluble globular structures (Ikehara et al., 2002). In the following paragraphs, the progress of investigation from the discovery of origin of genes to the GNC-SNS primitive genetic code hypothesis will be describe more precisely.

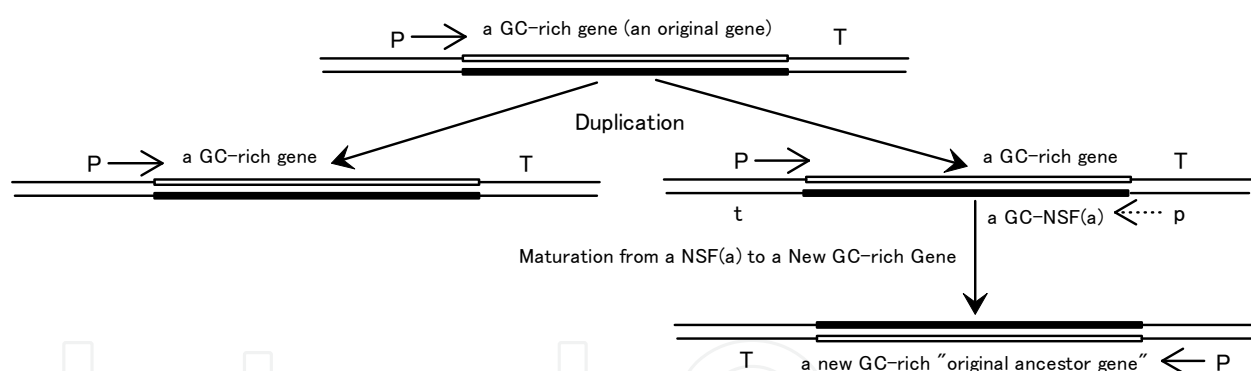


Fig. 5. GC-NSF(a) primitive gene hypothesis for creation of "original ancestor genes" under the universal genetic code. The hypothesis predicts that new "original ancestor genes" originate from nonstop frames on antisense strands of GC-rich genes (GC-NSF(a)s)

Firstly, we found that base compositions at the three codon positions of the GC-NSF(a) were similar to SNS. Actually, hypothetical polypeptide chains encoded by only SNS code, not containing A and U at the first and third codon positions, satisfied the six conditions, suggesting that polypeptides encoded by SNS code could be folded into water-soluble globular structures at a high probability (Figure 6 (A)). This indicates that SNS code has enough ability encoding proteins with definite-levels of catalytic activities. At this point, I provided SNS hypothesis on the origin of the genetic code about fifteen years ago (Ikehara & Yoshida, 1998).

But, the SNS code composed of 16 codons and 10 amino acids must be too complex to prepare as the first genetic code from the beginning. So, I further searched for which code

was more primitive one than SNS by using the four more essential conditions which acidic amino acid and basic amino acid compositions were excluded from the six conditions described above. From the results, it was found that [GADV]-proteins encoded by GNC codons well satisfied the four structural conditions, when roughly equal amounts of [GADV]-amino acids were contained in the proteins (Figure 6 (B)). Where [GADV] represents four amino acids of Gly, Ala, Asp and Val, and square bracket ([]) was used to discriminate amino acids, especially G and A which are described by one-letter symbols of amino acids, from nucleic acid bases, G and A. It means that even the [GADV]-polypeptide chains with a quite simple amino acid composition could be folded into water-soluble structures at a high probability.

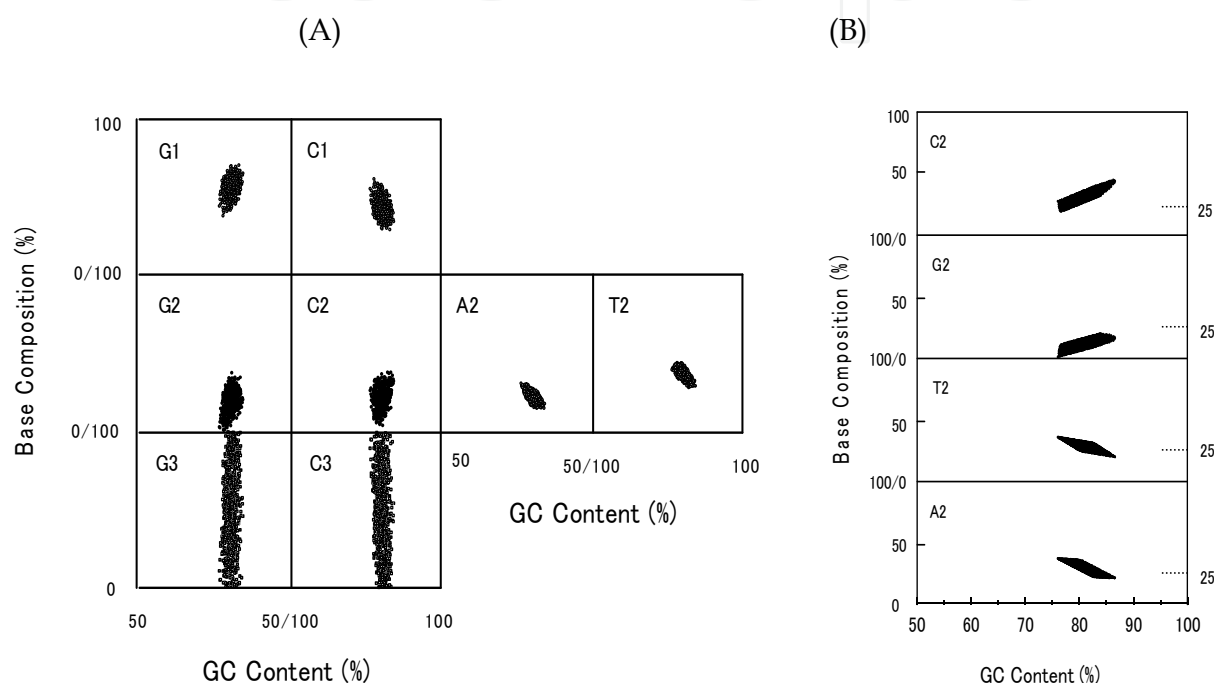


Fig. 6. (A) Dot plot analysis of SNS genetic code. Dots concentrated in the respective boxes indicate that the six conditions (hydropathy, α -helix, β -sheet and turn/coil formabilities, and acidic and basic amino acid contents) were satisfied. It means that polypeptide chains encoded by SNS code could be folded into water-soluble globular structures when bases are contained in the respective rates at three codon positions. (B) Dot plot analysis of GNC code

On the other hand, other codes encoding four amino acids, which were picked out from the columns or rows in the universal genetic code table, did not satisfy the four structural conditions, except for GNG code, which is a modified form of the GNC code (Ikehara et al, 2002). Moreover, it was also confirmed that genetic code composed of three amino acids lined in universal genetic code table did not satisfy the four conditions for protein structure formation, suggesting that the GNC code would be used as the most primeval genetic code on the primitive earth (Ikehara et al, 2002). Then, I concluded that SNS primitive genetic code evolved from the GNC primeval genetic code by C and G introductions at the first and the third codon positions, respectively (Figure 7 (A)).

Dots concentrated in the respective boxes of Figure 6 (B) indicate that the four conditions (hydropathy, α -helix, β -sheet and turn/coil formabilities) were satisfied. It means that polypeptide chains encoded by GNC code could be folded into water-soluble globular

structures when four bases are contained in the respective rates at the second codon position.

Thus, I provided GNC-SNS hypothesis as the origin of the genetic code about ten years ago (Ikehara et al., 2002), suggesting that the universal genetic code originated from GNC code through SNS code as capturing new codons up and down in the genetic code table (Figure 7 (B)).

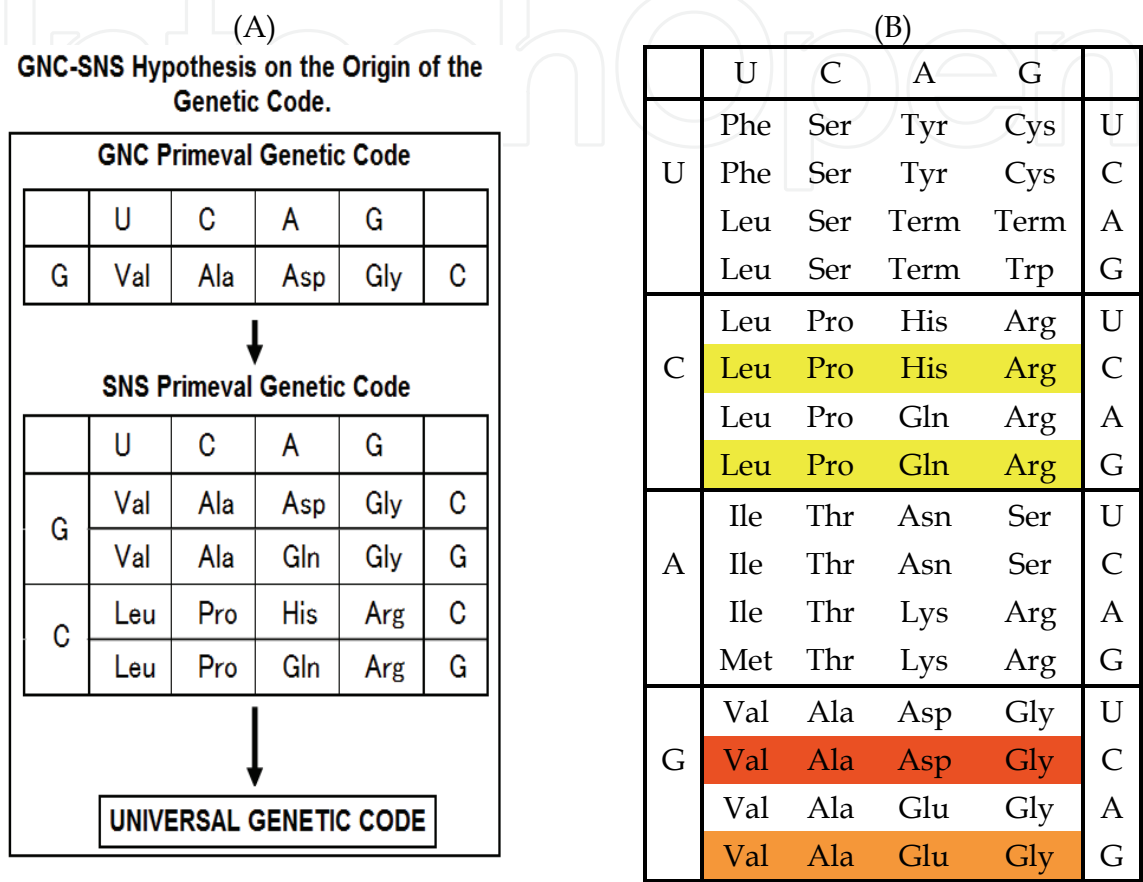


Fig. 7. GNC-SNS hypothesis on the origin and evolutionary pathway of the genetic code. (A) In the hypothesis, it is supposed that the universal genetic code originated from GNC primeval genetic code through SNS primitive genetic code. Elucidation of the most primitive GNC code made it possible to propose as GADV hypothesis on the origin of life. (B) Alternative representation of the origin and evolutionary pathway of the genetic code. The universal genetic code originated from GNC primeval genetic code (red row), successively followed by capturing codons of GNG (orange row), and CNS (yellow rows), resulting in formation of SNS code. Therefore, it is considered that the universal genetic code evolved from GNC code through the introduction of rest rows up and down

Due to the evolutionary process of the genetic code, amino acids with similar chemical/physical properties have been arranged in the same column at a high probability (Table 2). Consequently, replacements between two amino acids located in the same column have been permitted at a high probability and the robustness of the genetic code has been generated. Now I believe that the GNC code had stepped up its structure to the SNS primitive genetic code encoding ten amino acids with 16 SNS codons via GNS code (8 codons and 5 amino acids). After that, the SNS code evolved into the universal genetic code,

which encodes 20 amino acids and three stop signals with 64 codons (Ikehara & Yoshida, 1998; Ikehara et al., 2002). The GNC-SNS primitive genetic code hypothesis represents that the universal genetic code (NNN: $4 \times 4 \times 4 = 4^3 = 64$ codons), which is both formally and substantially triplet code, originated from formally triplet but substantially singlet GNC code ($1 \times 4 \times 1 = 4^1 = 4$ codons) encoding four [GADV]-amino acids, through formally triplet but substantially doublet SNS code ($2 \times 4 \times 2 = 4^2 = 16$ codons) encoding 10 amino acids (Figure 7) (Ikehara, 2009).

Evolutionary process of the genetic code from GNC code, encoding four amino acids with quite different chemical/physical properties, to the universal genetic code through SNS code arranged amino acids with similar chemical and physical properties in the same columns and with largely different properties in the same rows at high probabilities (Table 2). So, it is considered that the robustness of the genetic code originated from the evolutionary process of the genetic code as suggested by the GNC-SNS primitive genetic code hypothesis. The discussion on the robustness of the genetic code is consistent with the results of permissible amino acid substitutions, which were observed between two homologous proteins, as given in Figures 2 and 3. As described below, the finding of the GNC-SNS primitive genetic code hypothesis led to the ideas on protein 0th-order structures and on the origin of life as GADV hypothesis or [GADV]-protein world hypothesis (Ikehara, 2005; Ikehara, 2009).

4. The universal genetic code and protein 0th-order structure

Discussion on protein structure formation usually begins with primary structure or amino acid sequence of a protein, not with amino acid composition. In Stryer's textbook "Biochemistry" (Berg et al, 2002), it is described that the information needed to specify the catalytically active structure of ribonuclease is contained in its amino acid sequence. The studies on folding of polypeptide chains, which were mainly carried out with small-sized proteins, have established the generality of this central principle of biochemistry: sequence specifies conformation. One of the reasons may rely on the facts that one-dimensional base sequences on DNA or genes encode amino acid sequences or primary structure of proteins.

On the other hand, I happened to use amino acid composition for investigation of protein structure formability, the six or four conditions as described above. The utilization gave interesting results and conclusions, such as GC-NSF(a) hypothesis on creation of the first family genes and GNC-SNS primitive genetic code hypothesis as described in the previous Sections 3. During the investigation on the origin of the genetic code, I have noticed the significance of specific amino acid compositions satisfying four (hydropathy and α -helix, β -sheet and turn propensities) or six (hydropathy and α -helix, β -sheet and turn propensities plus acidic and basic amino acid compositions) conditions for folding polypeptide chains into water-soluble globular structures. The conditions were obtained as the respective average values plus/minus standard deviations of presently existing water-soluble globular proteins from seven micro-organisms carrying the genomes with widely distributed GC contents. Structure formability of one protein is the same as other proteins randomly assembled in the same amino acid composition. This means that every protein synthesized by random peptide bond formation among amino acids in the specific amino acid composition could be similarly folded into water-soluble globular structures, but into different structures, since the proteins have the same amino acid composition but different sequences from each other.

The most important point for creation of entirely new proteins encoded by the first family genes is to form water-soluble globular structure through random synthesis among amino acids in a protein 0th-order structure, because a quite large number of possible catalytic sites for an organic compound could appear on the surface of one globular protein. The number of possible catalytic sites can be estimated from combinations of amino acids locating on the protein surface as about several hundred points. I have named such a specific amino acid composition favorable for protein structure formation as protein 0th-order structure (Ikehara, 2009), for example, the compositions containing roughly equal amounts of four [GADV]-amino acids (Gly [G], Ala [A], Asp [D] and Val [V]) and ten amino acids ([GADV]-amino acids plus Glu [E], Leu [L], Pro [P], His [H], Gln [Q] and Arg [R]) encoded by GNC and SNS codes, as [GADV]- or GNC- and SNS-protein 0th-order structures, respectively. This means that the protein 0th-order structures are secretly written in the universal genetic code table (Figure 7 (B)).

Origins of genes and proteins: Genetic code plays a central role in connecting genetic function with catalytic function in the fundamental life system, as described above (Figure 4). Under the GNC code, the first genes must be composed of base sequences carrying only GNC codons, which were produced by random phosphodiester bond formation among GNC codons. Subsequently, the first double-stranded (GNC)_n gene would be created by complementary strand synthesis against the single-stranded (GNC)_n gene.

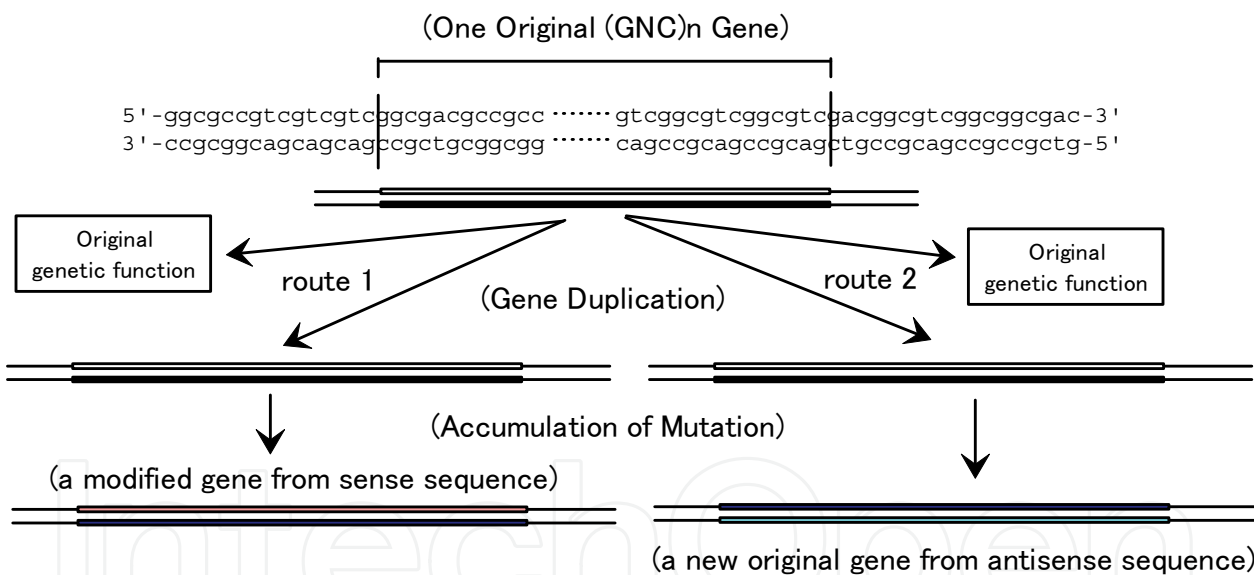


Fig. 8. Two routes for producing new genes. Once one original double-stranded (GNC)_n gene was produced, new genes were easily produced by using two base sequences (one is from sense sequence and the other is from antisense sequence) of the original gene or through two routes. From route 1, new genes could be produced as modified genes of the original gene or homologous genes in a gene family and from route 2, new genes could be created as “entirely new genes” or the first family genes

Creation of the first double-stranded (GNC)_n gene following establishment of the GNC primeval genetic code became the most important points leading to the emergence of life, since the invention of double-stranded genes made it possible for the first time to transmit genetic information from parents to progenies and to evolve it through accumulation of base substitutions and selection of more effective genetic sequences (Ikehara, 2009).

Base compositions at three codon positions on sense strands of $(\text{GNC})_n$ genes are substantially same as those on anti-sense strands, due to the self-complementary structure of the double-stranded $(\text{GNC})_n$ genes. Thus, it is easily supposed that, after creation of the first double-stranded $(\text{GNC})_n$ gene, GNC codon sequences on anti-sense strands could be utilized as a field for creation of entirely new functional genes encoding the first ancestor proteins in homologous protein families, since GNC codon sequences on antisense strands are quite different from those on sense strands, as can be actually regarded as random arrangement of GNC codons. In addition, $(\text{GNC})_n$ sequences on antisense strands must encode [GADV]-proteins satisfying the four conditions for producing water-soluble globular proteins at a high probability (Ikehara, 2002) (Figure. 6 (B)). Also new genetic information could be created from duplicated sense sequences, as proposed by Ohno (1970). But, the duplicated sense sequences could be utilized only for encoding homologous proteins in a family (route 1). Contrary to that, one of two antisense sequences obtained after gene duplication could give a field for production of the protein, which is quite different from all proteins existed before (route 2) (Figure 8) (Ikehara, 2009).

As seen in Figure 6 (B), [GADV]-proteins must have similar rigidity to extant proteins, when [GADV]-proteins contain less and more amounts of glycine and alanine than one quarter, respectively. Therefore, it is supposed that [GADV]-proteins, which were produced on the primitive earth in the absence of any genetic function or before creation of the first gene, were more flexible than the presently existing proteins, since the proteins should contain flexible turn/coil forming amino acid, glycine, more than rigid α -helix forming amino acid, alanine. The reason is that glycine would be pre-biotically synthesized more easily and accumulated on the primitive earth more than alanine. Therefore, [GADV]-proteins produced on the primitive earth must be more flexible than extant proteins recognizing usually one organic compound with high catalytic activities and high specificities. The flexible [GADV]-proteins would inevitably have only quite low catalytic activities. Even the low activities of the firstly appeared [GADV]-proteins would have been effective for leading to creation of the first genetic code, the first gene and the first life on the primitive earth. That is because the existence of [GADV]-proteins having the low catalytic activity must be important to develop new metabolic pathway on the primitive earth without any genetic information.

Formation of flexible but inefficient [GADV]-proteins was also essential to create newly-born proteins or the first family proteins even after the first double-stranded $(\text{GNC})_n$ gene was produced, because the proteins, which were newly produced as ones with quite low enzymatic activities, could evolve to mature enzymes through accumulation of base substitutions and selection of more efficient enzymes with more rigid structures and higher specificities for one organic compound than before.

In fact, I believe that entirely new proteins have been created and selected from water-soluble globular proteins encoded by GC-NSF(a)s similar to $(\text{SNS})_n$ or SNS repeating sequences, even at present, when necessary. Initially, entirely new proteins could be produced by transcription from cryptic promoters and translation of anticodon sequences on GC-rich genes if the proteins had pre-requisite catalytic functions (Figure 5). The newly-born proteins composed of 20 kinds of amino acids would evolve to mature enzyme with more rigid structure and a high specificity for one specific-organic compound through accumulation of mutations and selection of efficient enzymatic activity as similarly as the case of [GADV]-proteins encoded by $(\text{GNC})_n$ anticodon sequences. I have now understood the important role of protein 0th-order structures or specific amino acid compositions in

creation of entirely new proteins or the first family proteins. As a matter of course, mechanisms for the creation of entirely new proteins intimately related to the creation of entirely new genes. These new concepts on the origins of the genetic code, proteins and genes led to the GADV hypothesis on the origin of life.

5. GNC primeval genetic code and origin of life

In this Section, I will describe briefly GADV hypothesis on the origin of life, since the hypothesis, which I have proposed, is intimately related to the origin of the genetic code or the GNC primeval genetic code.

RNA world hypothesis has been proposed as a key idea for solving the “chicken and egg dilemma” observed between genes and proteins or the origin of life and has been widely accepted by many investigators at the present time. While I have proposed a novel hypothesis on the origin of life as GADV hypothesis, suggesting that life originated from [GADV]-protein world, which was composed of [GADV]-proteins accumulated by pseudo-replication of the proteins in the absence of any genetic function (Ikehara, 2002; Ikehara, 2005, Ikehara, 2009). In the hypothesis, it is assumed that life emerged from the world through establishment of GNC primeval genetic code followed by formation of single-stranded and double-stranded (GNC)_n genes.

I believe that the most important point for solving the riddle on the origin of life would be to understand the origin and evolutionary processes of the fundamental life system, which is composed of genetic function, genetic code and catalytic function (Figure 4), not always to solve the “chicken and egg dilemma” observed between genes and protein, as considered in the RNA world hypothesis. Therefore, the GADV hypothesis would be far more rational to explain the origin of life than the RNA world hypothesis, because the former can easily explain formation processes of the fundamental life system composed of genes, the genetic code and proteins comprehensively as well as the “chicken and egg dilemma” (Ikehara, 2009). Contrary to that, the RNA hypothesis probably cannot explain the ways how the fundamental life system was created, because the hypothesis based on self-replication of RNA, which is carried out by polymerization of nucleotides one-by-one, cannot explain the origins of the genetic code and genes, which are composed of codons having triplet nucleotide sequences.

6. Robustness of the universal genetic code

Most genetic disorders are quite rare as causing the disorders at a ratio of only one person in every thousands or millions. The frequency of a genetic disorder caused by one-base substitution mainly relies on mutation rate. But, as given in Figures 2 and 3, in the cases of homologous microbial proteins belonging in the same protein family, many amino acid substitutions are observed without largely affecting protein function. The reasons are given as followings. The first one is because, utilization of many kinds of amino acids would be permissible in flexible regions of a protein at a high probability, such as turn/coil structures connecting two secondary structures and unstructured segments observed at C-terminal segment and/or at N-terminal segment at a high frequency, as can be seen in Figure 2. The second one could be attributed to the robustness of the universal genetic code, making it possible to use the same amino acids and different amino acids but with similar chemical and physical properties, when base substitutions occurred at the third and the first codon

positions, respectively. Therefore, the robustness of the genetic code could protect from destroy of protein's active state at a high probability, even if base substitutions occurred at the third and the first codon positions in genetic sequences and even when amino acid substitutions were introduced at the sites of secondary structures as α -helix and β -sheet structures. In contrast, base substitutions at the second codon positions would affect largely the protein functions, leading to the genetic disorders at a high probability, as shown in Figure 9.

According to the GNC-SNS primitive genetic code hypothesis, it is considered that the genetic code originated from GNC successively to SNS and finally to the universal genetic code as expanding the code up and down in the genetic code table as described in Section 3. From the evolutionary pathway of the genetic code, it can be understood that codons encoding amino acids with similar and with chemically different amino acids were arranged in columns and rows of the genetic code table, respectively. In other words, it is considered that the genetic code evolved as raising coding capacity to modulate the protein function, and as capturing new codons encoding new amino acids into vacant positions of the previous code table during evolutionary process. Therefore, the robustness of the genetic code could be generated from the origin and evolutionary processes of the genetic code, as described below.

1. Base substitution at the first codon position, but introducing no base change at the second position, does not destroy protein function at a high probability, since codons in the same column of the genetic code table code for amino acids with comparatively similar chemical/physical properties, because amino acids with the same color background are arranged in two and one columns out of four columns of hydrophacy and turn/coil tables, respectively. This can be also confirmed from the facts shown in Table 2.
2. Base substitution at the second codon position largely destroys protein function at a high probability, since codons located in the same row of the genetic code table encode amino acids with quite different chemical/physical properties (Table 2). Certainly, amino acids with the same color background are not observed on any row of four tables, except for one row having two termination codons in Table 2 (C). Amino acids with two different color backgrounds are arranged in eighteen out of 64 rows of the four tables of Table 2, otherwise amino acids in the same rows have three color backgrounds.
3. Base substitutions at the third codon position induce no amino acid replacement due to the degeneracy of the genetic code and substitutions between amino acids with similar chemical/physical properties, such as Phe-Leu, Asp-Glu, His-Gln and so on, are observed at a high probability.

Generally speaking, only base substitutions occurred at the second codon position, not at the first and third codon positions, induce substitutions between amino acids with largely different chemical and physical properties. The skillful location of codons in the genetic code table gives the genetic code robustness against base substitutions on genetic sequences, which is derived from the origin and evolutionary process of the genetic code, as suggested by the GNC-SNS primitive genetic code hypothesis (Ikehara et al., 2005).

7. The universal genetic code and genetic disorder

Genetic disorders are actually caused by base changes on autosomes and sex-chromosomes as X-chromosome, or on genomes in organelles as mitochondria. The genetic disorders are

classified by location of genetic elements, as autosomal, X-linked, Y-linked and mitochondrial. Now, it is known that many patients are suffered from genetic disorders induced by one-base substitutions on DNA. Several representative genetic disorders are described in Table 1. For simplicity, genetic diseases induced by deletions and insertions of genetic sequences are excluded from the Table. The number of genetic disorders would be reach to the total number of genes (about from twenty to thirty thousands in human), since almost all genes are essential for organisms to live.

Besides classification by locations of genetic changes, the disorders are also classified by forms of the genetic disease appearance into descendants, as dominant and recessive. Genetic disorders caused by mutation of DNA sequences on genomes encoding metabolic enzymes, which leads to reduction of enzyme activities, such as ADA (adenosine deaminase) deficiency and PKU (phenylketonurea), are generally inherited in recessive manners. Autosomal recessive genetic disorders are not appeared into their children, if either parent has two normal genes on two chromosomes, and the disorders are inherited at a 25% chance if both parents are carriers of the disorder. Contrary to that, Huntington's disease and neurofibromatosis caused by inheritance of the abnormal genes from either parent are inherited dominant manner. Therefore, each child has a 50% chance upon inheriting the genetic disorder, if just one parent has a dominant gene defect.

Genetic disorders caused by one-base substitutions are induced when base changes in genetic sequences went across a framework of the robust genetic code or when the base changes made proteins not to satisfy the conditions for formation of water-soluble globular structures, resulting in collapsing the protein structures. As I have discussed in this Chapter, many patients would be suffered from genetic disorders upon even one-amino acid replacement at a high probability, if one-base substitution occurred at the second codon positions. As can be seen in Figure 9, ornithine transcarbamoylase deficiency (OTCD) appears, when one amino acid is replaced to other amino acid encoded by codon having different base at the second codon position, more frequently than the replacement occurring between amino acids encoded by two codons having different bases at the first codon position.

This makes a remarkable contrast with the amino acid replacements observed between homologous proteins with similarly active catalytic function as given in Figures 2 and 3. Therefore, it suggests that it is important to repress base substitutions at the second codon position in genetic sequences in order to protect from genetic diseases. It is necessary to recognize bases at the second base position of codon to accomplish the purpose. As genetic sequences or genes are codon sequences not always mere nucleotide sequences, it would be possible to discriminate the bases at the second codon position from bases at the other two codon positions, based on the differential base compositions at the three base positions in codons. The reason is that it is already known that codons in genetic sequences encoding microbial proteins have specific base compositions at the three respective base positions. For example, guanine bases are generally observed more frequently at the first codon position than other three bases, whereas relatively equal amounts of four bases are contained at the second codon position of GC-rich genes (Ikehara, et al. 1996), although it is almost impossible to find out the strategy for protection of base substitutions at the second codon position at the present time. But, it would be important to recognize the facts described above, as the first step of discovery of the strategies for repression of base replacements at the second codon position in genetic sequences. New possible genetic treatment discovered will release human beings from genetic disorders in a future.

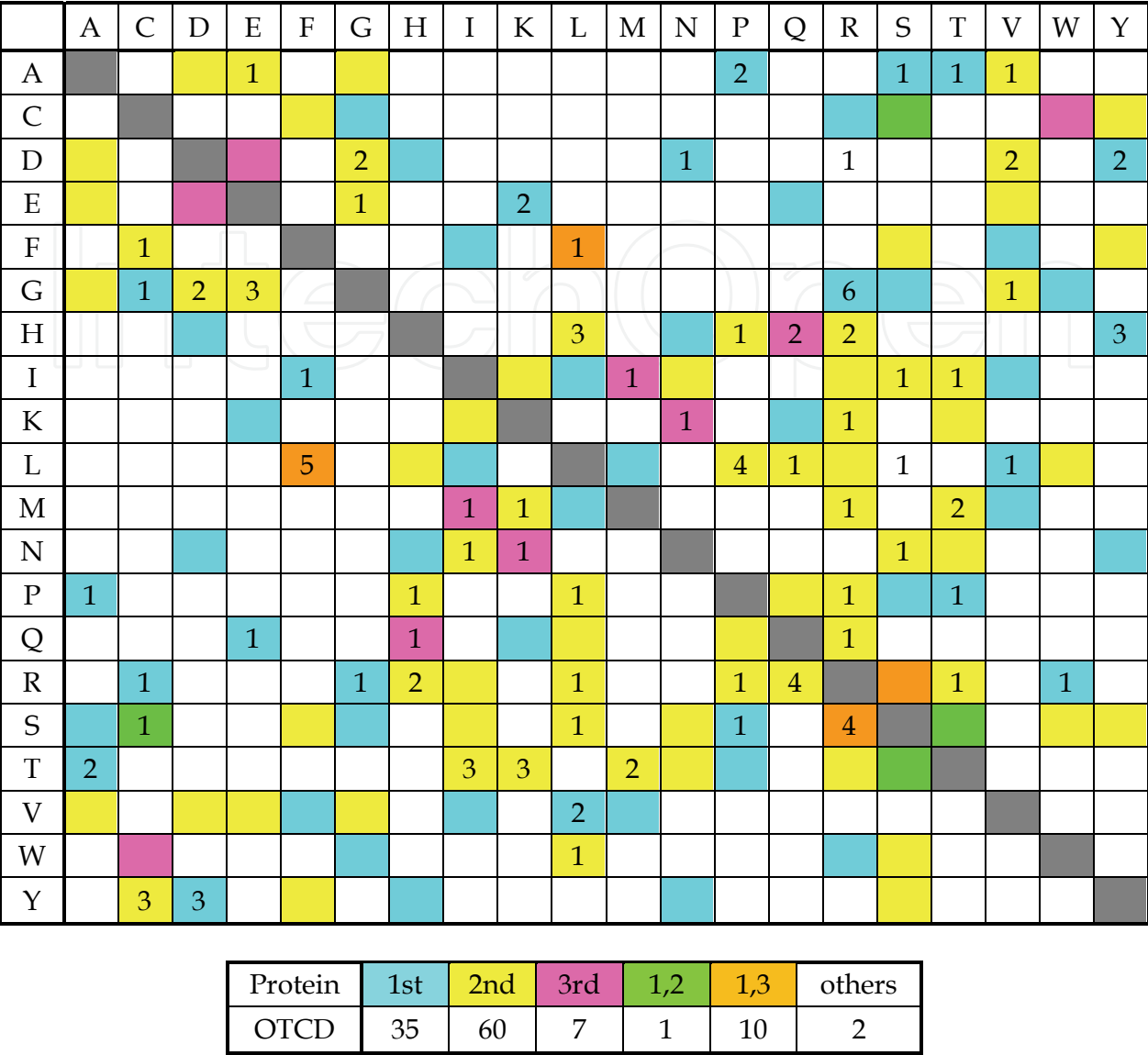


Fig. 9. Amino acid replacements observed in a genetic disorder, ornithine transcarbamoylase deficiency (OTCD). Letters written in the most left column and the top row indicate amino acids of normal ornithine transcarbamoylase described with one-letter symbols and those of mutated ornithine transcarbamoylase causing OTCD. Blue, yellow and red boxes indicate amino acid substitutions caused by base changes at the first, the second and the third codon positions, respectively. Green, orange and white boxes indicate amino acid replacements induced by base substitutions at the first or the second codon position, at the first or the third codon position and other base substitutions, respectively. Color box representation is the same as Figure 3. Data of the amino acid replacements observed in OTCD were obtained from Natural Variants in Protein Knowledgebase (UniProKB) at the address of <http://www.uniprot.org/uniprot/P00480>

8. Conclusion

The genetic disorders upon one-base substitutions in genes encoding amino acid sequences of proteins are induced by the base substitutions at the second codon position more

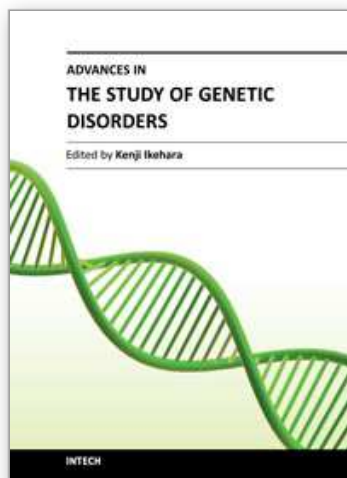
frequently than those at the first codon position. The fact intimately relates to the robustness of the genetic code, which is derived from the origin and evolutionary process of the genetic code. According to the GNC-SNS primitive genetic code hypothesis, which I have proposed, it is considered that the universal genetic code originated from GNC code through SNS code as expanding the code up and down in the genetic code table. Due to the origin and evolutionary process of the genetic code, amino acids with similar chemical and physical properties have been located in the same columns. The arrangement of amino acids in the genetic code table makes it possible to repress induction of genetic disorders at a low rate, because one-base substitutions at the first codon position do not largely affect protein functions at a high probability. I would like to say that it is important to understand correctly the main cause inducing the genetic disorders as the first step for protection of the diseases, and that the recognition will release human beings from many genetic disorders someday.

9. Acknowledgements

I am grateful to Dr. Tadashi Oishi (Narasaho College) for the encouragement of our research on GNC-SNS hypothesis on the genetic code and GADV hypothesis on the origin of life.

10. References

- Berg JM, Tymoczko JL, & Stryer L. (2002) Biochemistry 5th ed. New York: W. H. Freeman and Company.
- Ikehara, K. (2002) Origins of gene, genetic code, protein and life: comprehensive view of life system from a GNC-SNS primitive genetic code hypothesis. *J. Biosci.* 27, 165-186.
- Ikehara, K. (2005) Possible steps to the emergence of life: The [GADV]-protein world hypothesis. *Chem. Record*, 5, 107-118.
- Ikehara, K. (2009) Pseudo-replication of [GADV]-proteins and origin of life. *Int. J. Mol. Sci., (International Journal of Molecular Sciences)* Vol. 10, No. 4, 1525-1537.
- Ikehara, K., Amada, F., Yoshida, S., Mikata, Y., & Tanaka, A. (1996) A possible origin of newly-born bacterial genes: significance of GC-rich nonstop frame on antisense strand. *Nucl. Acids Res.*, 24, 4249-4255.
- Ikehara, K., Omori, Y., Arai, R. & Hirose, A. (2002) A novel theory on the origin of the genetic code: a GNC-SNS hypothesis. *J. Mol. Evol.*, 54, 530-538.
- Ikehara, K., & Yoshida, Y. (1998) SNS hypothesis on the origin of the genetic code. *Viva Origino*, 26, 301-310.
- Ohno, S. (1970) *Evolution by Gene Duplication*, Springer: Heidelberg, Germany.



Advances in the Study of Genetic Disorders

Edited by Dr. Kenji Ikehara

ISBN 978-953-307-305-7

Hard cover, 472 pages

Publisher InTech

Published online 21, November, 2011

Published in print edition November, 2011

The studies on genetic disorders have been rapidly advancing in recent years as to be able to understand the reasons why genetic disorders are caused. The first Section of this volume provides readers with background and several methodologies for understanding genetic disorders. Genetic defects, diagnoses and treatments of the respective unifactorial and multifactorial genetic disorders are reviewed in the second and third Sections. Certainly, it is quite difficult or almost impossible to cure a genetic disorder fundamentally at the present time. However, our knowledge of genetic functions has rapidly accumulated since the double-stranded structure of DNA was discovered by Watson and Crick in 1956. Therefore, nowadays it is possible to understand the reasons why genetic disorders are caused. It is probable that the knowledge of genetic disorders described in this book will lead to the discovery of an epoch of new medical treatment and relieve human beings from the genetic disorders of the future.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Kenji Ikehara (2011). Origin of the Genetic Code and Genetic Disorder, *Advances in the Study of Genetic Disorders*, Dr. Kenji Ikehara (Ed.), ISBN: 978-953-307-305-7, InTech, Available from:
<http://www.intechopen.com/books/advances-in-the-study-of-genetic-disorders/origin-of-the-genetic-code-and-genetic-disorder>

INTECH
open science | open minds

InTech Europe

University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China

Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2011 The Author(s). Licensee IntechOpen. This is an open access article distributed under the terms of the [Creative Commons Attribution 3.0 License](https://creativecommons.org/licenses/by/3.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

IntechOpen

IntechOpen