Degeneracy of the Genetic Code has Played an Important Role in Evolution of Organisms

Kenji Ikehara*

G&L Kyosei Institute, Hikaridai 1-7, Seika-cho, Kyoto 619-0237, Japan
International Institute for Advanced Studies, Kizugawadai 9-3, Kizugawa, Kyoto 619-0225, Japan

Received: May 28, 2016; Accepted: July 15, 2016; Published: July 25, 2016

***Corresponding author:** Kenji Ikehara, G&L Kyosei Institute, Hikaridai 1-7, Seika-cho, Kyoto 619-0237, Japan; Fellow, International Institute for Advanced Studies, Kizugawadai 9-3, Kizugawa, Kyoto 619-0225, Japan; Emeritus professor, Nara Women's University, Kita-uoya-higashi-machi, Nara 630-8506, Japan, Tel: +81-774-73-4478; Fax: +81-774-73-4478; E-mail: ikehara@cc.nara-wu.ac.jp

Abstract

The genetic code is degenerate mainly at the third codon position. However, the reason of the degeneracy still remains unknown. On the other hand, we have proposed GC-NSF(a) hypothesis, assuming that entirely new gene has emerged from non-stop frame on antisense codon sequence of GC-rich gene (GC-NSF(a)). Amino acid sequence of a protein encoded by GC-NSF(a) widely varies without any change of amino acid sequence of a protein produced from the corresponding gene on sense strand. Therefore, various protein sequences with an amino acid sequence can be repeatedly examined, if a required function is found on a surface of the protein from the GC-NSF(a). This indicates that the degeneracy of the genetic code has largely contributed to produce entirely new protein from GC-NSF(a). Thus, entirely new proteins having a required function could be produced effectively. The degeneracy of the genetic code made it possible for organisms to prosper on Earth. Organisms, which did not use a degenerate genetic code, would extinguish from this planet. This is one significant point of the genetic code.

Keywords: Protein 0th-order Structure; Degeneracy of the Genetic Code; Origin of Entirely New Gene

Introduction

Genetic information transferred from DNA to mRNA by transcription is transmitted to protein through translation on ribosome. During the translation, the universal or standard genetic code written with triplet base sequence is used, which specifies a kind of amino acid, like as that GUC, CAC, AAG and UUC correspond to Val, His, Lys and Phe, respectively. As well known, the genetic code is degenerate mainly at the third codon position, for example GGN, GCN and GUN, 4 codons each, code for Gly, Ala and Val, respectively (Figure 1). However, it is totally unknown about the reason why the genetic code is degenerate.

The genetic code occupies a core position connecting genetic function with catalytic function in the fundamental life system (Figure 2). This means that the genetic code is not only a simple representation of triplet base sequence with an amino acid, but also a key element when one wants to make clear formation process of the fundamental life system, which is composed of gene (genetic function), genetic code and protein (catalytic

function), leading to solving the riddle of the origin of life.

In this Short Communication, I describe that the degeneracy of the genetic code plays an important role in producing entirely new (EntNew) gene and the corresponding EntNew protein from nonstop frame on anti-sense strand of GC-rich gene (GC-NSF(a)), according to the GC-NSF(a) hypothesis, which I have proposed [1-3] (Figure 3). That would be the significance of the genetic code.

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

Figure 1: Degeneracy of the genetic code. Almost all extant organisms on the Earth are using so called universal or standard genetic code, which degenerates mainly at the third codon position of the code.

Figure 2: Position of genetic code in fundamental life system. Genetic code occupies a core position connecting genetic function composed of DNA and mRNA with catalytic function carried out protein and metabolism in the life system.

Theories proposed for new gene formation

Organisms on the Earth have evolved gradually and steadily, and sometimes abruptly, over several billion years. The evolution of lives should have tightly coupled to acquisition of new function required to adapt to new environment.

Two main routes for creation of new genes have been proposed so far. One is gene duplication theory proposed by Ohno [4], predicting that after duplication of a gene, one duplicate may acquire a new adaptive function, while the other duplicate retains the original function. The second is exon-shuffling theory proposed by Gilbert, et al. [5], assuming that new functional gene is created as a new combination of exons shuffled among two or more genes.

However, both of the theories do not explain the most fundamental problem on the creation of EntNew gene or the first ancestor gene in a family, from which various descendant genes could be derived. This means that the most important and key problem about formation of a gene family remains unsolved.

On the contrary, I have proposed GC-NSF(a) hypothesis on creation of EntNew gene [1-3]. The hypothesis assumes that EntNew gene, which is totally different from any previously existing gene, is created from GC-NSF(a) after gene duplication (Figure 3). I would like to emphasize here that GC-NSF(a) hypothesis is not a purely theoretical idea but one attained by database analysis of microbial genes and proteins, which have been obtained by experiments.

The reason why EntNew gene can be produced from GC-NSF(a)

The reason why EntNew genes can be produced according to the GC-NSF(a) hypothesis is as follows.

1. Protein encoded by GC-NSF(a) satisfies the six conditions for formation of water-soluble globular protein, which were obtained as average ranges (average value $\pm$ standard deviation) of six properties (hydropathy, α -helix, β -sheet and turn/coil formabilities, and acidic amino acid and basic amino acid compositions) of water-soluble

globular proteins encoded by genomes of seven extant microorganisms carrying chromosomal DNA with widely different GC content [6]. This indicates that a polypeptide chain produced from GC-NSF(a) would be folded into water-soluble globular structure at a high probability.

2. The reason why protein encoded by GC-NSF(a) well satisfies the six conditions is partly because GC-NSF(a) is similar to SNS repeating codon sequence or $(\text{SNS})_n$, which also satisfies the six conditions [7]. S and N mean G or C and any one of four bases.
3. GC-NSF(a) codes for a protein, which has completely different amino acid sequence from any previously existing protein but has a similar amino acid composition to those of proteins encoded by GC-rich genome (Figure 4). The protein encoded by GC-NSF(a) has a nearly random amino acid sequence in protein 0th-order structure. Protein 0th-order structure is a specific amino acid composition, in which even random joining of amino acids produces water-soluble globular protein at a high probability. One of such kind of protein 0th-order structures is GC-NSF(a)-encoding amino acid composition similar to SNS-encoding amino acid composition (Figure 4). Other protein 0th-order structures are SNS-encoding ten amino acids ([GADV]-amino acids plus Glu[E], Leu[L], Pro[P], His [H], Gln [Q] and Arg [R]) and GNC-encoding four [GADV]-amino acids. [GADV] means four amino acids; Gly[G], Ala[A], Asp[D] and Val[V].
4. As well known, sequence diversity of even a small protein composed of 100 amino acids is extraordinary large, as reaching to about $20^{100} \sim 10^{130}$. The tremendously large diversity makes it possible to produce protein without any homology with previously existing proteins or an EntNew protein from a GC-NSF(a), because the probability which an amino acid sequence of a protein arisen from a GC-NSF(a) coincides with any one of previously existing proteins is just about nil, as $\sim 1/10^{130}$. Of course, sequence diversity of a protein encoded by GC-NSF(a) is much smaller than $\sim 10^{130}$, since amino acid composition of the protein is extremely deviate from that of a protein containing about equal amount of 20 amino acids. However, the diversity is larger than $\sim 10^{100}$, the still extraordinary large diversity value of a protein encoded by $(\text{SNS})_{100}$ similar to codon sequence of GC-NSF(a) (Figure 4). Note that SNS codes for 10 amino acids.
5. In addition, the degeneracy at the third codon position increases the probability that EntNew protein is produced from GC-NSF(a), because the EntNew protein could be generated without any change of amino acid sequence encoded by a GC-rich gene on sense strand. The base replacement at the third codon position of GC-rich gene corresponds to the first codon position of GC-NSF(a) (Figure 5). If a required function were detected on a surface of the protein encoded by one of various GC-NSF(a)s, the protein could be newly born and successively evolve to mature protein (8).

Therefore, degeneration of the genetic code is quite important for production of EntNew protein; because every amino acid sequence appeared upon base replacement could be examined until a required catalytic activity is found.

Discussion

Base at the third codon position on a sense strand can be replaced widely without any change of amino acid sequence, because of degenerate of the genetic code at the position. On the other hand, this causes to change amino acid sequence encoded by GC-NSF(a) at a high probability, since the change of base at the first codon position induces amino acid replacement of a protein encoded by GC-NSF(a). Nevertheless, amino acid composition, not amino acid sequence, of the protein does not largely deviate from the protein 0th-order structure. The reason is because GC-

mutation pressure acting on GC-rich gene of a microorganism carrying GC-rich genome remains to keep the composition nearly in SNS-encoding amino acid composition, which is also one of the protein 0th-order structures (Figure 4).

Thus, it is considered that EntNew protein having a required function could be produced effectively owing to the degeneracy of the genetic code, and that organisms on this planet, which have used the degenerate genetic code, have prospered, although it might be as a result. Inversely saying, organisms, which did not use a degenerate genetic code, would extinguish from the Earth, even if the organisms once emerged on this planet. The genetic code of organisms prospering on this planet must be degenerate, for whatever reason for the degeneracy.

Then, from when did the genetic code degenerate? According to my scenario, GNC-SNS primitive genetic code hypothesis [6], on the origin and evolution of the genetic code, it can be supposed that the degeneracy started at the evolutionary step from the first genetic code GNC to the next code GNS, which encodes five amino acids, GNC-encoding [GADV]-amino acids plus Glu [E]. The GNS code might trigger to evolve it to SNS code and to produce effectively EntNew gene/protein from antisense codon sequence ((SNS)_n sequence) of (SNS)_n gene, as similarly as EntNew Gene/protein emerged from GC-NSF(a) (Figure 3).

Acknowledgements

I am very grateful to Dr. Tadashi Oishi (G&L Kyosei Institute, Emeritus professor of Nara Women's University) for encouragement throughout my research on GADV hypothesis on the origin of life.

References

- 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(21):4249-4255.
- Ikehara K. Simulation of gene evolution: evidence for GC-NSF(a) hypothesis on the origin of genes. *Viva Origino*, 2003;31(3):201-215.
- Ikehara K. Mechanisms for creation of "original ancestor genes". *J. Biol. Macromol.* 2005;5(2):21-30.
- Ohno S, "Evolution by Gene Duplication": 1970; Heiderberg, Springer.
- Gilbert W, de Souza SJ, Long M. Origins of genes. *Proc. Natl. Acad. Sci. USA* . 1977;94(15):7698-7703.
- 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(4):530-538. doi:10.1007/s00239-001-0053-6
- Ikehara K, Yoshida S. SNS hypothesis on the origin of the genetic code. *Viva Origino*, 1996; 26: 301-310.
- Ikehara K. Protein ordered sequences are formed by random joining of amino acids in protein 0th-order structure, followed by evolutionary process. *Orig. Life Evol. Biosph.* 2014;44(4):279-281. doi:10.1007/s11084-014-9384-3

Figure 4: Base compositions of GC-NSF(a) at three codon positions. Base compositions at three codon positions of both GC-NSF(a) and GC-rich gene are similar to each other and can be approximated to SNS. Base compositions at the three codon positions were obtained as average values of seven genes of *Pseudomonas aeruginosa* carrying GC-rich genome.

Amino acid sequence encoded by a GC-rich gene											
5'	A	H	P	E	T	G	K	3'	(Sense Strand)		
	GC	NC	AY	CC	NG	AR	AC	NG	GA	AR	
3'	CG	NG	UR	GG	NC	UY	UG	NC	NU	UY	5'
		Y	M	W	F	C	S	F			
		R	V	R	L	R	P	L			
		S		(R)		S	T				
		G		G		G	A				

Figure 5: Possible amino acid sequences encoded by a GC-NSF(a). For example, 1,536 (4x2x3x2x4x4x2) amino acid sequences can be encoded by only 7 codons of GC-NSF(a), while amino acid sequence encoded by a GC-rich gene is retained unchanged due to the degeneracy of the genetic code at the third codon position, as shown above.