Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 5088–5093, April 1998
Evolution

The role of robustness and changeability on the origin and evolution of genetic codes

TETSUYA MAESHIRO*† AND MASAYUKI KIMURA‡

*Department 6, ATR Human Information Processing Research Laboratories, 2–2 Hikaridai, Seika, Soraku, Kyoto, 619-0237 Japan; and ‡School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, 923-1292 Japan

Edited by Philip W. Anderson, Princeton University, Princeton, NJ, and approved February 2, 1998 (received for review September 22, 1997)

ABSTRACT We propose that an essential factor on the origin of genetic codes is a balanced accomplishment of robustness and changeability, two antithetical, but fundamental, properties for the survival and evolution of organisms. These measures are defined as the intrinsic properties of genetic codes. An evaluation of these properties explains the structural regularity of genetic codes, estimates the order of codon reassignment in deviant codes, and predicts the most probable deviant codes that exist. The enumeration of genetic codes that could have evolved from the standard genetic code under the selection pressure on robustness and changeability strongly limits the freedom of codon reassignments. The codon reassignments of all currently known deviant genetic codes belong to this predicted evolutionary path, and they generally give the highest improvements on robustness and changeability.

We propose that requests for both robustness and changeability have a strong influence on the origin of the standard genetic code (SGC) (Table 1) and its evolution to deviant codes (Table 2). These are paradoxical requests, and whereas the robustness is related to the survivability of organisms, the changeability is related to their evolvability. The investigation indicates that this is a reasonable possibility. The robustness is defined by two properties: the μ-robustness, which is the unalterability of phenotypes caused by a single base mutation of codons, where the phenotypes denote any of 20 amino acids and the stop codon; and the s-robustness, which is the robustness against nonsense mutations. The changeability is the alterability of phenotypes by a single base mutation of codons. These measures are intrinsic properties of genetic codes.

The elucidation of an increasing number of deviant codes (Table 2), where some codons are reassigned to different phenotypes, suggests that SGC is their ancestor (1). No general theory, however, explains the structural regularity of SGC and why it has evolved to many deviant codes. Current hypotheses, such as distance minimization of the polarity of amino acids (2–4), coevolution of amino acids and the genetic code (6), and maximum resistance against single base mutations (7), explain only partially the structure of SGC, and fail to explain the origin of deviant codes, which have occurred independently a number of times at least in ciliates (8). On the other hand, the biased codon usage was proposed as a mechanism to originate the deviant codes (1). Under a strong GC (or AT) pressure, only the codon whose third base is G/C (or A/U) would be used to code phenotypes assigned with multiple codons. Unused codons were free to change without affecting the functionality of organisms, originating deviances in the code. This did not, however, explain why deviant codes had appeared.

Robustness and Changeability of Genetic Codes

The genetic code is a coding table between 64 codons and 21 phenotypes. Theoretically, 21 phenotypes are assignable to 64 codons to minimally reflect the mutations in a DNA sequence on amino acid sequences, to increase the robustness against the mutations. Genetic codes with high robustness imply a low probability of change in amino acid sequences, but for a fixed mutation rate, a high reflection of mutations is advantageous for exploring proteins with new functions and for following environmental variations. Because necessary changes are unpredictable, a high average changeability between all phenotype pairs is advantageous.

Graph Visualization and DNA Mutation Model. The robustness and changeability of genetic codes are calculated based on their graph representation (Fig. 1). Some simplifications are made to specify the DNA mutation mechanism against which the genetic codes should be robust and changeable. First, DNA substitution models used in phylogenetic methods (9) are unused; for example, DNA substitution rates varying among lineages, because phylogenetic analysis treats the DNA sequences that are the result of a repetitive process of change in the DNA sequence and subsequent selection. We assume that the robustness and changeability of the genetic codes are related solely to the mutation of the DNA sequence because no environmental changes can be predicted. Therefore, mutations observed in pseudo genes are most appropriate. Second, nucleotide substitution is assumed to be the most influential mutation mechanism. Consequently, insertions and deletions, which are about 10 times less frequent than the nucleotide substitution (10), are ignored. Finally, unbiased codon usage is assumed because of the wide intraspecific variations in the codon usage among genetic systems using the same genetic code. For example, the GC content on the silent base varies between 2% and 59% among species using the deviant code MNe, and the variation increases with any increase in the size of the available DNA sequence data (11). We model the bias of the mutation rate between transition pairs (GC to AT and AT to GC), which is probably the primary cause of variations in the GC content in DNA (12). The existence of some unpredictable factors, such as tRNA abundance (13), is another reason to ignore the codon usage. Indeed, such a bias is easily modeled, as the graph structure is unmodified.

Initially, we use an even mutation rate because the concept of the robustness and changeability of genetic codes becomes clearer. The transition-transversion bias and GC-AT bias affect neither the graph structure nor the concept. Then, biased mutation rates explain the detailed structures of the genetic codes, and reinforce our explanation on the possible origin of the genetic codes.

μ-Robustness. Let s_i be the set of codons in node i of a graph, and n_i be the number of codons in s_i, denoted as the size of s_i. Then, the μ-robustness r_i of node i, denoted as individual μ-robustness r_i, is

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

This paper was submitted directly (Track II) to the *Proceedings* office.
Abbreviation: SGC, standard genetic code.
†To whom reprint requests should be addressed.

Table 1. Standard genetic code

1st base	2nd base								3rd base
	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA*	*stop*	UGA*	*stop*	A
	UUG$^+$		UCG		UAG*	*stop*	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG$^+$		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG$^+$	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

The codons marked with + are chain-initiator or initiation codons, and stop codons are chain-terminating codons.

$$r_i = \frac{1}{9n_i}\sum_{j=1}^{n_i} u_{ij}, \qquad \mathbf{[1]}$$

where u_{ij} is the number of single base mutants of the j-th codon in s_i, which belong to the same set s_i, e.g., UUU and UUC of node F (Phe) are each other's single base mutant on their third base, and $9n_i$ is the total number of single base mutants generated by n_i codons, because each of the codons' three bases generates three single base mutants. Here, r_i denotes the probability to keep coding the phenotype i corresponding to node i against single base mutations in the set of n_i codons. For example, $r_F = 1/(9{\cdot}2) \times (1 + 1) = 1/9$. The μ-robustness $\bar{r}$ of genetic codes, one of two kinds of robustness, is the average of all r_i,

$$\bar{r} = \frac{1}{v}\sum_i r_i, \qquad \mathbf{[2]}$$

where v is the total number of nodes in the graph. Hence, to maximize the individual robustness r_i is to assign a set of n_i codons to node i to maximize the number of single base mutant pairs m_i. A set of p codons coding the same phenotype that differ only in a single base is called a p-column set because it corresponds to a column made up of p unit cubes as shown in Fig. 2, where $p = 1, \ldots, 4$, and 4 is the number of genotypes, i.e., A, C, G, and U. The number of single base mutant pairs of a p-column set is p times $(p - 1)$, the maximum number of pairs among p codons. In the case of $4 < n_i \leq 8$, two column sets, i.e., a four-column set and an $(n_i - 4)$-column set, give the maximum number of single base mutant pairs.

Proposition. To assign q phenotypes, $16 \leq q \leq 64$, a genetic code has the maximum μ-robustness, if and only if, the codons assigned to each phenotype constitute p-column sets, $1 \leq p \leq 4$, so that q p-column sets constitute 16 four-column sets in the cubic representation of 64 codons (11).

To understand the proposition, let $r_i(n_i)$ be the maximum individual μ-robustness of node i with size n_i, $1 \leq n_i \leq 64$. Assigning more than four codons to any node or phenotype decreases the μ-robustness $\bar{r}$, because the average number of assigned codons per phenotype is less than 4, and the incremental value of $r_i(n_i)$ for $n_i \geq 5$ is smaller than that for $1 \leq n_i \leq 4$, if n_i codons constitute a single p column set. Note that the values of $r_i(p)$ of node i configured in the p-column set are $r_i(1) = 0$, $r_i(2) = 1/9$, $r_i(3) = 2/9$, and $r_i(4) = 3/9$. If the size of the p-column set of node i increases to $p + 1 \leq 4$, then the size of the p'-column set of some other node j decreases to $(p' - 1) \geq 1$. The increase of $r_i(p)$ to $r_i(p + 1) = r_i(p) + 1/9$ compensates for the decrease of $r_j(p')$ to $r_j(p' - 1) = r_j(p') - 1/9$, keeping the value of μ-robustness $\bar{r}$.

Changeability. The changeability of a genetic code is the alterability of phenotypes caused by a single base mutation of codons. It measures the average of the transition probabilities along the shortest paths between all of the pairs of phenotypes in the graph representation of the code, because the shortest paths between the nodes practically determine the transition probabilities, and consequently, the changeability of the code.

Given two nodes i with size n_i and j with size n_j connected with an edge, let m_{ij} and m_{ji} be the number of single base mutant pairs from nodes i to j and j to i, respectively, where clearly $m_{ij} = m_{ji}$. Then, the transition probability from node i to j is $m_{ij}/9n_i$, as m_{ij} of a total of $9n_i$ single base mutants belongs to node j. Similarly, the transition probability from node j to i is $m_{ij}/9n_j$. Then, their average, denoted as path width ρ_{ij}, is

$$\rho_{ij} = (m_{ij}/9n_i + m_{ij}/9n_j)/2. \qquad \mathbf{[3]}$$

For a pair of nodes not directly connected with an edge, for example, node j linking nodes i and k, the path width between i and k is the average of the transition probabilities of paths $i \rightarrow j \rightarrow k$ and $k \rightarrow j \rightarrow i$, given respectively by $m_{ij}/9n_i \times m_{jk}/9n_j$ and $m_{kj}/9n_k \times m_{ji}/9n_j$. When multiple paths exist between two nodes, only the path widths of the shortest paths are summed. When two or more nodes correspond to a phenotype, such as Ser, the shortest paths from all relevant nodes are considered. The paths between the nodes of amino acids linked by a stop node, denoted as interrupted paths, are removed in the calculation of the path widths, because these paths correspond to the nonsense mutations that result in the synthesis of shorter proteins, and most of them have no biological activity (18). Note that all of the shortest paths are considered if node i or j is a stop node. For example, in Fig. 1, the shortest paths between nodes Y and W are Y-stop-W, Y-C-W, and Y-S4-W, but the first one, an interrupted path, is removed and $\rho_{YW} = \rho_{YC} \times \rho_{CW} + \rho_{YS4} \times \rho_{S4W}$.

Then, the changeability $\bar{\rho}$ of a genetic code is the average of the path widths between all of the pairs of phenotypes, which is 210 for 21 phenotypes

$$\bar{\rho} = \frac{1}{210}\sum_i \sum_j \rho_{ij}, \quad i \neq j. \qquad \mathbf{[4]}$$

We also measure the connectivity of phenotype i with all phenotypes in a genetic code, denoted as individual changeability ρ_i,

$$\rho_i = \sum_j \rho_{ij},\ i \neq j. \qquad \mathbf{[5]}$$

s-Robustness. The s-robustness, which measures the robustness against nonsense mutations, considers the interrupted paths that were excluded from the calculation of the changeability $\bar{\rho}$ because of the deleterious consequences of nonsense mutations.

Let ρ_{ij} be the path width between nodes i and j as defined in Eq. **3**, where interrupted paths are excluded. Now, let ρ'_{ij} be the path width of the interrupted paths between nodes i and j. We define the s-robustness φ as

$$\varphi = \frac{\Sigma_i \Sigma_j \rho'_{ij}}{\Sigma_i \Sigma_j \rho_{ij} + \Sigma_i \Sigma_j \rho'_{ij}}, \qquad \mathbf{[6]}$$

which is the ratio of the total interrupted path width to the total path width. A smaller value of φ implies a better s-robustness, because the s-robustness measures the probability of nonsense mutations during missense mutations, which involve the mutation of a codon assigned to one amino acid into a codon assigned to a different amino acid.

Table 2. Assignments of deviant codons

Representative genetic system	Code	Changes from SGC		Initiation codons	
		Codon	Phenotype		
Mitochondrial yeasts	MYe	UGA	stop ⇒ Trp	AUG	1
		AUA	Ile ⇒ Met		
		CUN	Leu ⇒ Thr		
Mitochondrial platyhelminths	MPl	UGA	stop ⇒ Trp	AUG	1
		AAA	Lys ⇒ Asn		
		AGR	Arg ⇒ Ser		
		UAA	stop ⇒ Tyr		
Mitochondrial nematoda	MNe	UGA	stop ⇒ Trp	AUN UUG GUG	6
arthropoda		AGR	Arg ⇒ Ser		
mollusca		AUA	Ile ⇒ Met		
Mitochondrial echinodermata	MEc	UGA	stop ⇒ Trp	AUG	1
		AAA	Lys ⇒ Asn		
		AGR	Arg ⇒ Ser		
Mitochondrial tunicata	MTu	UGA	stop ⇒ Trp	AUG	1
		AUA	Ile ⇒ Met		
		AGR	Arg ⇒ Gly		
Mitochondrial vertebrata	MVe	UGA	stop ⇒ Trp	AUN GUG	5
		AUA	Ile ⇒ Met		
		AGR	Arg ⇒ stop		
Mitochondrial euascomycetes	MEu	UGA	stop ⇒ Trp	AUN NUG UUA	8
Nuclear mycoplasma	CMy	UGA	stop ⇒ Trp	AUN NUG UUA	8
Nuclear euplotes	CEu	UGA	stop ⇒ Cys	AUG	1
Nuclear acetabularia	CAc	UAR	stop ⇒ Gln	AUG	1
Nuclear blepharisma	CBl	UAG	stop ⇒ Gln	AUG	1
Nuclear candida	CCa	CUG	Leu ⇒ Ser	AUG CUG	2
Nuclear bacterial	CBa	—	—	AUN NUG	7

N denotes any of A, U, G, and C, and R denotes A and G. The values in the initiation codons indicate the number of known initiation codons. The codon reassignments of each deviant code are arranged from top to bottom in the estimated order of reassignments. Compiled from http://www3.ncbi.nlm.nih.gov/htbin-post/Taxonomy/wprintgc?mode=c.

Relevance of Robustness and Changeability. The robustness (μ-robustness and s-robustness) and changeability of genetic codes become relevant when the DNA sequence changes, particularly through replication. These measures are related with the survivability and evolvability (adaptability) of species. With a high μ-robustness, the probability to conserve the protein sequence and its functionality is high. On the other hand, a high changeability gives larger variations of amino acid sequences after replications.

Suppose that some fitness function is given. Note that the fitness is evaluated at the amino acid sequence level, which is translated from the DNA sequence using the genetic code. For the same DNA sequence and mutation rate, offsprings replicated from organisms with genetic codes having a high changeability will have greater variations in the amino acid sequences than those with genetic codes having a high μ-robustness. Such variations would be advantageous under a variable environment. However, the offsprings of an organism with a high fitness are more likely to have a high fitness, if they are similar to their parents, because mutations are introduced randomly. In other words, although a genetic code with high changeability is easier to originate offsprings with different amino acid sequences, the probability is low that the offsprings also will have a high fitness.

Therefore, in a population where half of the inhabitants have a genetic code with high μ-robustness and the other half have a high changeability, the genetic code with the high μ-robustness likely will predominate the population if evaluated with the same fitness function, the same mutation rate, and without changes in their genetic codes. This has been verified through computer simulation.

The s-robustness is related to both the robustness and changeability, and measures the probability of nonsense mutations when an amino acid mutates into another amino acid. Genetic codes with a high s-robustness (low φ) allow mutations between amino acids with a low probability of nonsense mutations when two or more single base mutations are necessary.

Understanding Why SGC Has a Highly Regular Structure

Biased Selection Pressure on Robustness. The presence of selection pressure on μ-robustness is evident, as codon sets assigned to 20 amino acids constitute column sets to maximize their individual μ-robustness with a consequent high μ-robustness of SGC. It is notable that initiation codons also constitute a column set. Leu, Arg, Ser, and the stop codons violate the proposition above, decreasing the μ-robustness $\bar{r}$ by 17% relative to the optimal μ-robustness, counterbalanced by a typical increase in the changeability $\bar{\rho}$ by 14%. Genetic codes with the optimal μ-robustness have a very low value of changeability and vice versa, reflecting the contradictory nature of the requests for robustness and changeability and the difficulty of improving them simultaneously.

The optimality of the μ-robustness of SGC is 82.9% compared with a theoretical genetic code with the maximum μ-robustness, and the changeability is 42.6% relative to a theoretical genetic code with the maximum changeability. These values suggest a biased selection pressure on the robustness in the formation of SGC, because the population of organisms having genetic codes with a high robustness is probably more advantageous for survival purposes than a population with a high changeability, even in a highly variable environment. If many genetic codes were to compete during the establishment process of the standard, it would be possible for the code with a high μ-robustness to predominate, which is SGC.

Nodes of Leu, Arg, and Ser Increase the Changeability. Six codons are assigned to Leu, Arg, and Ser, and they effectively increase the changeability of SGC, because more codons assigned to a phenotype or a node increase the connectivity with other nodes. Hence, the nodes of Leu, Arg, and Ser function as dispatchers to facilitate the transitions between

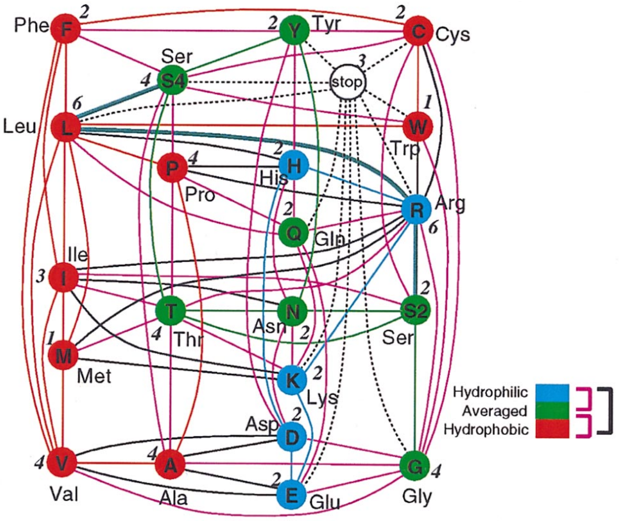

FIG. 1. Graph representation of SGC with the polarity of amino acids. The letters in the nodes are one-letter abbreviations of amino acids, except for S2 and S4. Each node in the graph is defined as a set of codons that code the same phenotype, where the codons in the set can change to any other in the same set through successive transitions with a single base mutation connected by a line called an edge, if the single base mutant of a codon in one node belongs to the other node; then, 20 phenotypes except for Ser correspond in one to 20 nodes of the graph. By the definition of the node, the set of codons coding Ser is divided into two nodes S2 = {AGU, AGC} and S4 = {UCU, UCC, UCA, UCG}. Red denotes a hydrophilic, green denotes an averaged (neutral), and blue denotes a hydrophobic amino acid. Connections between the amino acids are classified with colored edges, e.g., a purple edge represents a connection between a neutral and either a hydrophobic or hydrophilic amino acid. The numbers give the sizes of the nodes. The number of vertices with size 1, 2; size 2, 10; size 3, 2; size 4, 6; size 5, 0; and size 6, 2.

nodes of amino acids with similar polarities, as Leu is hydrophobic, Arg is hydrophilic, and Ser is averaged or neutral. The polarity is the strongest physico-chemical constraint on the protein functionality (19), which explains a balanced distribution of the polarities of the three dispatchers. It is interesting that their four nodes (A, R, S2, and S4) are directly connected, possibly to facilitate the transitions between amino acids with different polarities. Furthermore, Ser is unique for its two split nodes, supposedly to further facilitate the transitions, especially between amino acids with different polarities, because of its averaged polarity. It is notable that Ser is the only phenotype with multiple nodes in all deviant codes.

Very Delicate Role of the Stop Node. Although the presence of three stop codons suggests their average importance, as the average number of codons assigned to 21 phenotypes is $64/21 = 3.05$, the request for a low individual μ-robustness can be verified in the configuration of three stop codons not in the column set. In fact, the individual μ-robustness of this configuration is decreased by 33% from that of the three-column set, and its individual changeability is increased by 20% from the total individual changeability including the interrupted paths of Ile constituting the three-column set, for example.

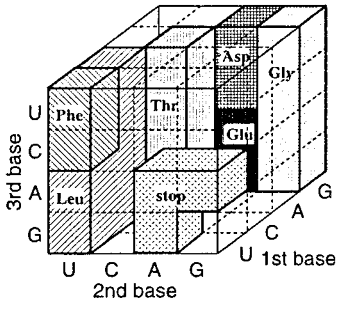

FIG. 2. Three-dimensional representation of SGC. The codons of some phenotypes are shown. Note that the set of three codons assigned to stop is not subject to the proposition in the text and does not constitute a three-column set.

The s-robustness typically is improved by 5% over a hypothetical code with stop codons constituting a three-column set. A low individual μ-robustness and a high individual changeability of the stop node are supposed to be consequences of a drastic loss of the protein functionality by nonsense mutations and the importance of stop codons to terminate the protein synthesis, resulting in a balanced accomplishment of the μ-robustness and the recoverability from nonsense mutations. However, this accomplishment is assumed to be very delicate, because stop codons are reassigned to amino acids in almost all deviant codes, which is an effective strategy for improving the s-robustness.

Absence of Nodes with Size 5. The model explains the absence of nodes constituting a five-column set in SGC and in 12 of 13 analyzed deviant codes. If a high μ-robustness is required, nodes should constitute column sets. For a high μ-robustness, the maximum node size should be 4 from the proposition. On the other hand, nodes with a size 6, whose individual μ-robustness is equal to that of a size 4, function as dispatchers to increase the changeability of the code. The individual μ-robustness of a five-column set is smaller than that of a four-column or six-column set, and any size increase of a node implies a size reduction of other nodes, because the total number of available codons is fixed to 64, indicating no advantage in assigning phenotypes or nodes with five-column sets. The node of a size 5 in the code CCa is not in the column set, probably to increase the individual changeability.

Optimality in the Evolution of Deviant Codes

Classification of Deviant Codes Based on Robustness and Changeability. The improved robustness and changeability of currently known deviant codes (Table 2) compared with those of SGC suggest SGC to be their evolutionary ancestor, where four types of selection pressure are identified (Fig. 3). The four types are (*i*) unbiased improvement on the robustness and changeability, further classified as improved μ-robustness, s-robustness, and changeability (codes MTu, MEu, CEu, and CMy) and improved s-robustness and changeability (codes CAc and CBl); (*ii*) biased improvement toward robustness (codes MNe, MEc, and MPl); and (*iii*) biased improvement toward changeability (codes MYe, MVe, and CCa).

Interpretation of Codon Reassignments. The possible reasons for the deviances from SGC (Table 2) are as follows. (*i*) An increase in the individual μ-robustness of newly assigned phenotypes, e.g., Trp (all mitochondrial codes and code CMy) and Met (codes MYe, MNe, MTu, and MVe). (*ii*) An increase in the s-robustness, such as Gln (codes CAc and CBl). (*iii*) Changes of the dispatchers, such as Ser, Thr, and Gly. The balanced distribution of polarities of the three dispatchers in SGC becomes concentrated to averaged or neutral amino acids in deviant codes MTu, MNe, MEc, MPl, MYe, and CCa. The node sizes of those dispatchers with extreme polarities, Arg and Leu, are reduced, increasing the node size or creating alternative nodes of Ser, Thr, or Gly, three amino acids with average polarities. This is possibly to ease the transitions between amino acids with different polarities, which is another way of improving the alterability of the phenotypes. (*iv*) An increase in the recoverability from nonsense mutations by splitting the node of stop codons, and increasing the individual changeability of stop codons (code MVe).

It is notable that the set of initiation codons constitutes two perpendicular column sets to maximize the individual μ-robustness in deviant codes with more than four initiation codons; the codes MEu, MNe, MVe, CMy, and CBa. Differing from the stop codons, the μ-robustness is exclusively required for the initiation codons and no recoverability is needed, as they correspond to amino acids in the middle of genes.

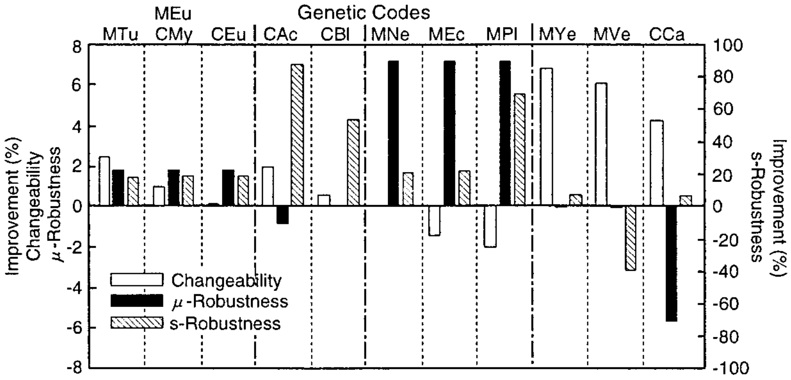

FIG. 3. Relative variations in changeability, robustness, and nonsense robustness of deviant codes compared with those of SGC. The values are given by $(v' - v)/v$ [$(v - v')/v$ for s-robustness], where v' is the value of a deviant code, and v is a value of SGC. The deviant codes are referred to by the abbreviations listed in Table 2, and are classified according to the manner of improvement.

Estimation of Codon Reassignment Order. It is possible to estimate the order of codon reassignments in deviant codes. For example, the code MTu has three deviances (Table 2). Therefore, we generate all possible codes from SGC with one and two reassignments using the three deviances, and estimate the order so that the robustness and/or changeability are successively improved. It is assumed that one deviance is introduced at a time, and the requirement for robustness is stronger than that for changeability in the early evolutionary stage, as suggested by the biased selection pressure on the robustness in SGC. Then, the estimated orders of codon reassignments are uniquely determined as shown in Table 2. The results coincide with those obtained from a phylogenetic analysis (11).

Prediction of Deviant Codes. The possible presence of selection pressure on robustness and changeability enables the prediction of deviant codes that could have evolved from SGC (Fig. 4). Deviant codes are predicted by assuming that one deviance is introduced at a time, and each tRNA recognizes exactly one codon to model the change in the anticodon list. Therefore, deviances emerge either by the change in an amino acid associated with a tRNA, or by the appearance/disappearance of a tRNA. After the introduction of deviances, hypothetical codes coding 21 phenotypes are classified by the manner of improvement on the robustness and changeability. Those without any improvement are rejected. We denote the number of all hypothetical codes coding 21 phenotypes as the total number of codes. Hypothetical codes with two deviances are generated from the selected codes with one deviance. The process repeats for more deviances, simulating the evolution of deviant codes from SGC.

Generally, the codon reassignments found in deviant codes (Table 2) give the highest improvement on the robustness and/or changeability, indicating the nonrandomness of the origin of deviant codes. The size of the set of predicted codes relative to the total number of codes is small, and becomes smaller for more deviances, as the total number increases exponentially. For example, only two of 1,240 possible deviances improve unbiasedly the robustness and changeability of SGC, and both reassignments are found in deviant codes, where that with the highest improvement (UGA stop ⇒ Trp) is estimated as the first reassignment introduced in all mitochondrial deviant codes. The degree of optimality lowers when the changeability is improved. This is because of the global character of the changeability, contrary to the local nature of the μ-robustness, so there are many possible configurations to improve the changeability. Additional constraints, such as the intensification of the role of Ser as a dispatcher, increases the optimality and reduces the number of predicted deviant codes, where the optimality of CAU (Leu) → Ser in code CCa becomes fifth among 22 possible reassignments, and the increase on the recoverability from nonsense mutations, which is the reassignment UCU (Arg) → stop in code MVe, becomes the most optimal among four.

Some organisms use a genetic code that codes UGA as selenocysteine (Sec) (20), coding 22 phenotypes. In this genetic code, only the dispatchers, Arg, Leu, and Ser, violate the proposition. Compared with optimal genetic codes coding 22

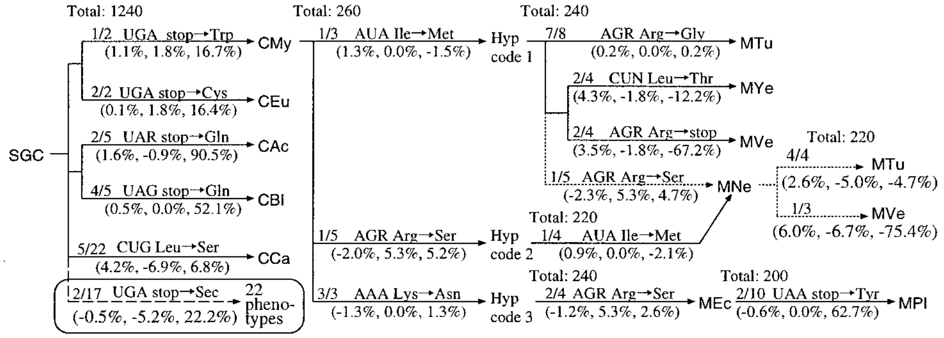

FIG. 4. Deviant codes predicted from SGC represented as an evolutionary tree. Only the codon reassignments found in known deviant codes are shown. The anticodon recognition pattern of existing mitochondrial codes (for examples, see refs. 5 and 23) was used to predict mitochondrial codes. The numbers labeled total indicate the total number of codes with 21 phenotypes. The edges are labeled with the number of predicted codes with a similar manner of improvement, the optimality of the reassignment, and the numbers in parentheses are successive improvement rates in the changeability, μ-robustness, and s-robustness. Note that a positive improvement in the s-robustness means a decrease in φ. Hyp code means hypothetical, unknown genetic codes. For example, of the total of 260 possible codes derived from the code CMy that code 21 phenotypes, 1/3 indicates that of the three possible reassignments that gives an improved changeability, the same μ-robustness, and a lower s-robustness relative to CMy, the highest improvement on changeability—1—is found in known deviant codes (1.3%, 0.0%, −1.5%). Dashed edges indicate possible but unlikely transitions, for example, MTu and MVe derived from MNe. Ser was probably introduced before Met in MNe, because reassignments of Gly (code MTu), Thr (code MYe), and stop (code MVe) improve mainly the changeability, contrary to the reassignment of Ser that decreases the changeability. Additional constraints besides the improved changeability on the reassignments Leu → Thr and Arg → stop from Hyp code 1 are respectively increased number of dispatchers with neutral polarity, and improved individual changeability of stop codons. The reassignment enclosed by a box generates a genetic code with 22 phenotypes, found in some organisms, where Sec is the additional amino acid.

phenotypes, the optimality of the μ-robustness is 84.3%, higher than that of SGC, and the changeability is 33.2%, lower than that of SGC. Furthermore, the reassignment UGA stop → Sec gives the second-highest improvement on the s-robustness and the lowest decrease on the changeability among 17 valid reassignments from our hypothesis, where the total number of possible reassignments is 62.

The use of the appropriate anticodon list of SGC (1) reduces the size of predicted sets and predicts the reassignment of multiple codons, for example, the reassignment of codons AGA and AGG to Ser, where both are recognized by anticodon UCU in mitochondria.

Influence of Transition-Transversion Biased Mutation Rate. Generally, the robustness (μ-robustness and s-robustness) of genetic codes increases and their changeability decreases for a transition mutation rate higher than the transversion. For a transition rate twice the transversion, the variations relative to even mutation rates are +16% in the μ-robustness, +9% in the s-robustness, and −5% in the changeability. On the other hand, the robustness decreases and the changeability increases for easier mutability of GC base pairs than AT. For a mutability of GC twice that of AT, the variations relative to even mutation rates are −2.4% in the μ-robustness, −12.8% in the s-robustness, and +0.6% in the changeability. These indicate a stronger influence of the mutation bias on robustness than changeability. The higher μ-robustness for a stronger transition is because of the grouping of codons with transition pairs (A-G and T(U)-C) in the same node when four codons differing on the third base are divided into two groups, such as CAU/CAC coding His and CAA/CAG coding Gln. The degeneracy on a single base is stated by the proposition for the maximum μ-robustness, and a transition-biased mutation rate gives a higher individual μ-robustness to keep coding the same phenotype, thereby increasing the μ-robustness of the genetic code. The biased mutation rate does not affect the highest individual changeability of the dispatchers, and the robustness and changeability of deviant codes are improved for any bias.

The spontaneous mutation of DNA replication, estimated from *Escherichia coli* with a defective error correction mechanism, seems to be highly frequent (24 times) in transition mutations, and has a higher mutability (1.8 times) of AT base pairs (14). An analysis on mammalian pseudogenes, which are probably free from selective constraints, gives different values, i.e., transitions 1.5–1.9 times more frequent and GC pairs 1.3–1.5 times easier to mutate (15–17).

It is possible that the mutation rate became less biased through evolution of the error correction mechanisms of the DNA replication, decreasing the robustness of the genetic code. Under this condition, deviant genetic codes with a higher robustness originated in species under conditions favorable for the deviation, and such genetic codes could have predominated in these species. This scenario could have happened because those codes with high robustness increase the survivability of the population over those codes with a high changeability.

Conclusions

The model provides a theoretical basis for understanding the central role of genetic codes, which point to the origin of life. Our work suggests that an important role of genetic codes is to determine how a change in a DNA sequence is reflected on amino acid sequences.

The present model accounts for the three essential properties of genetic codes, which are changeability, μ-robustness, and s-robustness, and should be positioned as the basis for more detailed analyses and models. Biased codon usage and influences of insertion and deletion are easily incorporated. The model quantitatively evaluates genetic codes and accurately predicts known deviant codes even without the appropriate anticodon list. For example, only three hypothetical codes improve the s-robustness and changeability without decrease in the μ-robustness if one tRNA recognizes each stop codon, and even with double recognition, as found in *Tetrahymena* (21), only five codes are possible. This explains the independent and multiple origin of deviant codes in ciliata (8), as selection pressure might act on these species to improve the s-robustness and changeability. Furthermore, some species of candida and ciliates have mitochondrial and nuclear deviant codes, for example, CCa and MYe, and CEu and MEu (data compiled from http://www3.ncbi.nlm.nih.gov/Taxonomy/tax.html), and similar improvements are found in both deviant genetic codes of the same species. This enables the prediction of nuclear or mitochondrial deviant codes for species with deviant codes in one of their genetic systems. The independence of the codon reassignment process found in ciliates is probably general. For example, the code MNe is found in nematoda and arthropoda, whose common ancestor is almost the origin of animals. If the codon reassignment were a rare event, all animal mitochondrial codes would have the deviances of MNe, but no such data is observed. Our hypothesis gives most probable codon reassignments, and it explains the presence of similar deviant genetic codes in phylogenetically distant species.

The concept of robustness and changeability offers a plausible explanation on the structure of SGC and its evolution to deviant codes. Although physico-chemical factors certainly influence the evolution of genetic codes, we propose that selection pressure on the robustness and changeability is also present. The choice of reassigned codons and newly assigned phenotypes in deviant codes seems to be nonrandom. If new deviant codes are discovered, they probably belong to the set of predicted codes. The assumption that the robustness and changeability are prerequisites for the survival and evolution of organisms is applicable to all aspects of evolution, for example, the disparity DNA replication hypothesis (5). The requirements for robustness and changeability are perhaps the single most universal aspect underlying the evolution of life.

We thank T. Keii and E. Tamiya for their helpful comments on the manuscript, T. Ikemura for valuable discussions, and K. Shimohara and K. Wada for their support.

1. Osawa, S., Jukes, T. H., Watanabe, K. & Muto, A. (1992) *Microbiol. Rev.* **56,** 229–264.
2. Wong, J. T. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 1083–1086.
3. Haig, D. & Hurst, L. D. (1991) *J. Mol. Evol.* **33,** 412–417.
4. Di Giulio, M., Capobianco, M. R. & Medugno, M. (1994) *J. Theor. Biol.* **168,** 43–51.
5. Furusawa, M. & Doi, H. (1992) *J. Theor. Biol.* **157,** 127–133.
6. Wong, J. T. (1976) *Proc. Natl. Acad. Sci. USA* **73,** 2336–2340.
7. Figureau, A. & Pouzet, M. (1984) *Origins Life* **14,** 570–588.
8. Tourancheau, A. B., Tsao, N., Klobutcher, L. A., Pearlman, R. E. & Adoutte, A. (1995) *EMBO J.* **14,** 3262–3267.
9. Huelsenbeck, J. P. & Rannala, B. (1997) *Science* **276,** 227–232.
10. Saitou, N. & Ueda, S. (1994) *Mol. Biol. Evol.* **11,** 504–512.
11. Maeshiro, T. (1997) Ph.D. thesis (School of Information Science, Japan Advanced Institute of Science and Technology).
12. Sueoka, N. (1993) *J. Mol. Evol.* **37,** 137–153.
13. Ikemura, T. (1981) *J. Mol. Biol.* **151,** 389–4009.
14. Schaaper, R. M. & Dunn, R. L. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 6220–6224.
15. Li, W.-H., Wu, C.-I. & Luo, C.-C. (1984) *J. Mol. Evol.* **21,** 58–71.
16. Bains, W. & Bains, J. (1987) *Mutat. Res.* **179,** 65–74.
17. Blake, R. D., Hess, S. T. & Nicholson-Tuell, J. (1992) *J. Mol. Evol.* **34,** 189–200.
18. Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A. & Weiner, A. M. (1986) *Molecular Biology of the Gene* (Benjamin-Cummings, Menlo Park, CA).
19. Lim, W. A. & Sauer, R. T. (1989) *Nature (London)* **339,** 31–36.
20. Zinoni, F., Birkmann, A., Stadtman, T. C. & Böck, A. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 4650–4654.
21. Hanyu, N., Kuchino, Y., Nishimura, S. & Beier, H. (1986) *EMBO J.* **5,** 1307–1311.