Biochemistry. In the article “The TRAP220 component of a thyroid hormone receptor-associated protein (TRAP) coactivator complex interacts directly with nuclear receptors in a ligand-dependent fashion” by Chao-Xing Yuan, Mitsuhiro Ito, Joseph D. Fondell, Zheng-Yuan Fu, and Robert G. Roeder, which appeared in number 14, July 7, 1998, of Proc. Natl. Acad. Sci. USA (95, 7939–7944), the authors wish to acknowledge an earlier paper that had escaped their attention. In an article entitled “Identification of RB18A, a 205 kDa new p53 regulatory protein which shares antigenic and functional properties with p53” [Drane, P., Barel, M., Balbo, M. & Frade, R. (1997) Oncogene 15, 3013–3024], Drane et al. report the identification of a human protein, RB18A, that interacts with several anti-p53 monoclonal antibodies, shows general DNA binding properties, and stimulates p53 binding to DNA. The cDNA-derived protein sequences of RB18A and TRAP220 are nearly identical, there being minor sequence variations and an extended N terminus on TRAP220. Apart from the effect of RB18A on p53 binding to DNA, the study of Drane et al. did not report any additional functions of RB18A and, in contrast to the findings with TRAP220, provided no indications of its presence within a larger multiprotein coactivator complex.

Biochemistry. In the article “Molecular cloning and characterization of a cellular phosphoprotein that interacts with a conserved C-terminal domain of adenovirus E1A involved in negative modulation of oncogenic transformation” by Ute Schaeper, Janice M. Boyd, Sulekha Verma, Erik Uhlmann, T. Subramanian, and G. Chinnadurai, which appeared in number 23, November 7, 1995 of Proc. Natl. Acad. Sci. USA (92, 10467–10471), we reported the sequences for human CtBP. Reexamination of the cDNA sequences revealed certain errors. These errors have been corrected in the GenBank database (accession no. U37408).

Evolution. In the article “The role of robustness and change-ability on the origin and evolution of genetic codes” by Tetsuya Maeshiro and Masayuki Kumura, which appeared in number 9, April 28, 1998, of Proc. Natl. Acad. Sci. USA (95, 5088–5093), the authors wish to note the following error in Table 2 related to the initiation codons of nuclear mycoplasma. They should read: AUG, GUG, and UUG [Dybvig, K. & Voelker, L. L. (1996) Annu. Rev. Microbiol. 50, 25–57]. Consequently, CMy (page 5091, last paragraph, right column) should be deleted in the text. These corrections do not change the results of the paper. A corrected Table 2 is shown below.

Table 2. Assignments of deviant codons

<table>
<thead>
<tr>
<th>Representative genetic system</th>
<th>Code</th>
<th>Changes from SGC</th>
<th>Phenotype</th>
<th>Initiation codons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mitochondrial yeasts</td>
<td>MYe</td>
<td>UGA stop ⇒ Trp</td>
<td>AUG</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AU A le ⇒ Met</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CUN Leu ⇒ Thr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitochondrial platyhelminths</td>
<td>MPI</td>
<td>UGA stop ⇒ Trp</td>
<td>AUG</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AAA Lys ⇒ Asn</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AGR Arg ⇒ Ser</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitochondrial nematoda</td>
<td>MNe</td>
<td>UGA stop ⇒ Trp</td>
<td>AUN UUG GUG</td>
<td>6</td>
</tr>
<tr>
<td>arthropoda mollusca</td>
<td></td>
<td>AGR Arg ⇒ Ser</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AAA Lys ⇒ Asn</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AUA Ile ⇒ Met</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitochondrial echinodermata</td>
<td>MEc</td>
<td>UGA stop ⇒ Trp</td>
<td>AUG</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AAA Lys ⇒ Asn</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AGR Arg ⇒ Ser</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitochondrial tunicata</td>
<td>MTu</td>
<td>UGA stop ⇒ Trp</td>
<td>AUG</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AUA Ile ⇒ Met</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AGR Arg ⇒ Gly</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitochondrial vertebrata</td>
<td>MVe</td>
<td>UGA stop ⇒ Trp</td>
<td>AUN GUG</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AUA Ile ⇒ Met</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AGR Arg ⇒ stop</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mitochondrial euascomycetes</td>
<td>MEu</td>
<td>UGA stop ⇒ Trp</td>
<td>AUN NUG UUA</td>
<td>8</td>
</tr>
<tr>
<td>Nuclear mycoplasma</td>
<td>CMy</td>
<td>UGA stop ⇒ Trp</td>
<td>AUG GUG UUG</td>
<td>3</td>
</tr>
<tr>
<td>Nuclear cuplotes</td>
<td>CEu</td>
<td>UGA stop ⇒ Cys</td>
<td>AUG</td>
<td>1</td>
</tr>
<tr>
<td>Nuclear acetalbularia</td>
<td>CAc</td>
<td>UAR stop ⇒ Gin</td>
<td>AUG</td>
<td>1</td>
</tr>
<tr>
<td>Nuclear blepharisma</td>
<td>CBl</td>
<td>UAG stop ⇒ Gin</td>
<td>AUG</td>
<td>1</td>
</tr>
<tr>
<td>Nuclear candida</td>
<td>CCa</td>
<td>CUG Leu ⇒ Ser</td>
<td>AUG CUG</td>
<td>2</td>
</tr>
<tr>
<td>Nuclear bacterial</td>
<td>CBa</td>
<td>—</td>
<td></td>
<td>AUN NUG</td>
</tr>
</tbody>
</table>

Plant Biology. In the article “Differential expression of two isopentenyl pyrophosphate isomerases and enhanced carotenoid accumulation in a unicellular chlorophyte” by Zairen Sun, Francis X. Cunningham, Jr., and Elisabeth Gantt, which appeared in number 19, September 15, 1998, of Proc. Natl. Acad. Sci. USA (95, 11482–11488), the following correction should be noted. In Fig. 3 the lightly shaded sequences, representing amino acid identity of four of five sequences, was inadvertently lost in the electronic submission process.

**FIG. 3.** Amino acid sequence alignment of IPP isomerases predicted by cDNAs of *Haematococcus pluvialis* (IPHp1 and IPHp2) and *Chlamydomonas reinhardtii* (IPICr1) are compared with sequences from the flowering plants *Clarkia brewerii* (IPICb2) (29) and *Arabidopsis thaliana* (IPIAt1) (30). The sequences were aligned using the program CLUSTALW (http://dot.imgen.bcm.tmc.edu:9331/multialign) and shaded using GENEDEC (http://www.concentric.net/~Ketchup/gddl.htm). The dark shading with white letters indicates amino acid identity for the aligned residue for all five proteins whereas the lightly shaded sequences represent amino acid identity for four of five sequences. The lines above the sequences indicate differences in the predicted amino acid sequences of IPHp1 and IPHp2. The inverted arrowhead (▼) above IPHp1 designates the location of a truncation from the N terminus. Residues required for catalytic activity (31) are marked by upright arrowheads (▲).
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

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 \( \mu \)-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 standard code (SGC) (Table 1) and its evolution to deviant codes (Table 2). However, such a bias is easily modeled, and the bias of the mutation rate between transition pairs (GC to AT and AT to GC) is probably the primary cause of variations 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.

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 not 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.

\( \mu \)-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 \( \mu \)-robustness \( r_i \) of node \( i \), denoted as individual \( \mu \)-robustness \( r_i \), is

\[ r_i = \frac{1}{n_i} \sum_{c \in s_i} \frac{1}{2} \left( 1 - \frac{1}{2^{c_3}} \right) \]

where \( c_3 \) is the third base of codon \( c \).
Table 1. Standard genetic code

<table>
<thead>
<tr>
<th>1st base</th>
<th>2nd base</th>
<th>3rd base</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>TTG</td>
<td>Stop</td>
</tr>
<tr>
<td>C</td>
<td>GUC</td>
<td>Val</td>
</tr>
<tr>
<td>G</td>
<td>AAC</td>
<td>Asn</td>
</tr>
<tr>
<td>A</td>
<td>UUC</td>
<td>Leu</td>
</tr>
<tr>
<td>AC</td>
<td>AGC</td>
<td>Ser</td>
</tr>
<tr>
<td>UU</td>
<td>UAG</td>
<td>Trp</td>
</tr>
<tr>
<td>UG+</td>
<td>AGC</td>
<td>Ser</td>
</tr>
</tbody>
</table>

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

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

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 \( n_i \) 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 corresponding to node \( i \) against single base mutations in the set of \( n_i \) codons. For example, \( r_{AC} = 1/(9 \times 2) = 1/18 \).

\[ \tilde{r} = \frac{1}{v} \sum_{i} r_i, \]  

where \( v \) is the total number of nodes in the graph. Hence, to maximize the individual robustness \( r_i \), 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 cubics as shown in Fig. 2, where \( p = 1, \ldots, 4 \), and 4 is the number of all phenotypes, i.e., U, C, G, and U. The number of single base mutant pairs of a \( p \)-column set is \( p^2 \) (\( p - 1 \)), 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 \( p \)-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 \( r_{ij}(n_i) \) be the maximum individual \( m \)-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 \( m \)-robustness \( \tilde{r} \), because the average number of assigned codons per phenotype is less than 4, and the incremental value of \( r_{ij}(n_i) \) 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 \( r(p) \) of node \( i \) configured in the \( p \)-column set are \( r(1) = 0, r(2) = 1/9, r(3) = 2/9, \) and \( r(4) = 3/9 \). 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 \( r(p) \) to \( r(p + 1) = r(p) + 1/9 \) compensates for the decrease of \( r(p') \) to \( r(p' - 1) = r(p') - 1/9 \), keeping the value of \( m \)-robustness \( \tilde{r} \).

**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 \( m_{ji}/9n_j \). Then, their average, denoted as path width \( p_{ij} \), is

\[ p_{ij} = (m_{ij}/9n_i + m_{ji}/9n_j)/2. \]  

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 \( j \rightarrow i \rightarrow k \), given respectively by \( m_{ij}/9n_i \times m_{jk}/9n_k \) and \( m_{ji}/9n_j \times m_{kj}/9n_k \). 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 \( p_{YW} = p_{YC} \times p_{CW} + p_{YS4} \times p_{S4W} \).

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

\[ \bar{p} = \frac{1}{210} \sum \sum p_{ij}, \quad i \neq j. \]  

We also measure the connectivity of phenotype \( i \) with \( j \) in a genetic code, denoted as individual changeability \( \rho_{ij} \),

\[ \rho_{ij} = \sum p_{ij}, \quad i \neq j. \]

**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 \( \bar{p} \) because of the deleterious consequences of nonsense mutations.

Let \( p_{ij} \) be the path width between nodes \( i \) and \( j \) as defined in Eq. 3, where interrupted paths are excluded. Now, let \( p_{ij} \) be the path width of the interrupted paths between nodes \( i \) and \( j \). We define the s-robustness \( \varphi \) as

\[ \varphi = \frac{\sum_{i} \sum_{j} p_{ij}}{\sum_{i} \sum_{j} p_{ij} + \sum_{i} \sum_{j} p'_{ij}}, \]  

which is the ratio of the total interrupted path width to the total path width. A smaller value of \( \varphi \) 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.
Relevance of Robustness and Changeability. The robustness (\(\mu\)-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 \(\mu\)-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 \(\mu\)-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 \(\mu\)-robustness and the other half have a high changeability, the genetic code with the high \(\mu\)-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 \(\varphi\)) 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 \(\mu\)-robustness is evident, as codon sets assigned to 20 amino acids constitute column sets to maximize their individual \(\mu\)-robustness with a consequent high \(\mu\)-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 \(\mu\)-robustness \(\tau\) by 17% relative to the optimal \(\mu\)-robustness, counterbalanced by a typical increase in the changeability \(\rho\) by 14%. Genetic codes with the optimal \(\mu\)-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 \(\mu\)-robustness of SGC is 82.9% compared with a theoretical genetic code with the maximum \(\mu\)-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 \(\mu\)-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
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.

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 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 MTe, MEu, CMe, and CMy) and improved s-robustness and changeability (codes CAc and CB1); (ii) biased improvement toward robustness (codes MNe, MEc, and MPI); and (iii) biased improvement toward changeability (codes MYe, MVe, and CCa).

**Interpretation of Codon Reassignments.** The possible reasons for the deviations 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 Glu (codes CAc and CB1). (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, MPI, 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 possible 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 MTeu, 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.
Estimation of Codon Reassignment Order. It is possible to estimate the order of codon reassessments in deviant codes. For example, the code MTu has three deviations (Table 2). Therefore, we generate all possible codes from SGC with one and two reassessments using the three deviations, and estimate the order so that the robustness and/or changeability are successively improved. It is assumed that one deviation 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 reassessments 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 deviation is introduced at a time, and each tRNA recognizes exactly one codon to model the change in the anticodon list. Therefore, deviations 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 deviations, 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 deviations are generated from the selected codes with one deviation. The process repeats for more deviations, simulating the evolution of deviant codes from SGC.

Generally, the codon reassessments 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 deviations, as the total number increases exponentially. For example, only two of 1,240 possible deviations 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 requirement for robustness and changeability are successively improved. It is assumed that one deviation 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 reassessments 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 deviation is introduced at a time, and each tRNA recognizes exactly one codon to model the change in the anticodon list. Therefore, deviations 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 deviations, 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 deviations are generated from the selected codes with one deviation. The process repeats for more deviations, simulating the evolution of deviant codes from SGC.

Generally, the codon reassessments 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 deviations, as the total number increases exponentially. For example, only two of 1,240 possible deviations 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 requirement for robustness and changeability are successively improved. It is assumed that one deviation 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 reassessments are uniquely determined as shown in Table 2. The results coincide with those obtained from a phylogenetic analysis (11).
phenotypes, the optimality of the $\mu$-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 $\rightarrow$ Sec gives the second-highest improvement on the $s$-robustness and the lowest decrease on the changeability among 17 valid realignments from our hypothesis, where the total number of possible realignments is 62.

The use of the appropriate anticodon list of SGC (1) reduces the size of predicted sets and predicts the realignment of multiple codons, for example, the realignment 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 ($\mu$-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 $\mu$-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 $\mu$-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 $\mu$-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 $\mu$-robustness, and a transition-biased mutation rate gives a higher individual $\mu$-robustness to keep coding the same phenotype, thereby increasing the $\mu$-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 1.5–1.3 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, $\mu$-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 $\mu$-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 ciliates (8), as selection pressure might act on these species to improve the $s$-robustness and changeability. Furthermore, some species of candida and ciliates have mitochon-ordial 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 deviations of MNe, but no such data is observed. Our hypothesis gives most probable codon realignments, 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.

---