Evolution of the genetic code; Evidence from serine codon use disparity in *Escherichia coli*

Masayori Inouye[a,1], Risa Takino[a], Yojiro Ishida[a], and Keiko Inouye[a]

[a]Department of Biochemistry, Center for Advanced Medicine and Biotechnology, Rutgers-Robert Wood Johnson Medical School, Piscataway, NJ 08854

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2019.

Contributed by Masayori Inouye, September 24, 2020 (sent for review July 16, 2020; reviewed by Milton H. Saier Jr and Dieter Söll)

Among the 20 amino acids, three of them—leucine (Leu), arginine (Arg), and serine (Ser)—are encoded by six different codons. In comparison, all of the other 17 amino acids are encoded by either 4, 3, 2, or 1 codon. Peculiarly, Ser is separated into two disparate Ser codon boxes, differing by at least two-base substitutions, in contrast to Leu and Arg, of which codons are mutually exchangeable by a single-base substitution. We propose that these two different Ser codons independently emerged during evolution. In this hypothesis, at the time of the origin of life there were only seven primordial amino acids: Valine (coded by GUX [X = U, C, A or G]), alanine (coded by GCX), aspartic acid (coded by GAY [Y = U or C]), glutamic acid (coded by GAZ [Z = A or G]), glycine (coded by GGX), Ser (coded by AGY), and Arg (coded by CGX and AGZ). All of these were derived from GGX for glycine by single-base substitutions. Later in evolution, another class of Ser codons, UCX, were derived from alanine codons, GCX, distinctly different from the other primordial Ser codon, AGY. From the analysis of the *Escherichia coli* genome, we find extensive disparities in the usage of these two Ser codons, as some genes use only AGY for Ser in their genes. In contrast, others use only UCX, pointing to distinct differences in their origins, consistent with our hypothesis.

serine codons | evolution | primitive amino acids | LUCA

More than half a century has passed since the genetic codons encoding all amino acids have been determined (1, 2). However, there are still many questions that remain unanswered. How did specific genetic codons evolve associations with individual amino acids and how were present universal genetic codons used in all living organisms on earth from bacteria to archaea, plants, and animals selected during evolution? It should be noted that there are variations to the universal nature of the genetic code most commonly found in organelles, such as in mitochondria and protozoan nuclei, but these differences are considered exceptions (3). Furthermore, the number of codons per amino acid is highly variable, from one to six. Of the 64 possible triplet codons consisting of four bases (guanine, adenine, uracil, and cytosine), 61 codons are assigned for 20 amino acids. Two amino acids are encoded by only one codon (methionine [Met] and tryptophan [Trp]), while the others are encoded by two (phenylalanine [Phe], tyrosine [Tyr], cysteine [Cys], histidine [His], glutamine [Gln], asparagine [Asn], lysine [Lys], aspartic acid [Asp], and glutamic acid [Glu]), three (isoleucine [Ile]), four (proline [Pro], threonine [Thr], valine [Val], alanine [Ala], and glycine [Gly]), or six codons (leucine [Leu], serine [Ser], and arginine [Arg]) (Table 1). The number of genetic codons for each amino acid does not generally reflect the abundance of the amino acids in living organisms. For example, although Leu, Ser, and Arg are encoded by six codons, their abundancies are not the highest in living organisms except for Leu. For instance, in *Escherichia coli* the abundancy ranks of Ser and Arg are no. 9 and no. 7, respectively.

Furthermore, among the three amino acids encoded by six codons—Leu, Arg, and Ser—the codons for Leu and Arg share at least one base in the first or the second position of the triple codons. For example, UUZ (Z = A or G) and CUX (X = U, C, A or G) for Leu, and CGX and AGZ for Arg, so that they can be mutually exchangeable by one-base substitutions between the synonymous codons, respectively. However, in the case of the Ser codons, AGY (Y = U or C) or UCX, at least two-base substitutions in both the first and the second positions are required between AGY and UCX to retain their capabilities to encode Ser. One-base substitutions either at the first or the second position (boldface/italics letters) result in encoding different amino acids. For example, AGY to ***G***GY (Gly), to ***C***GY (Arg), to ***U***GY (Cys), to A***A***Y (Asn), to A***C***Y (Thr), and to A***U***Y (Ile). These present six different possible changes to six different amino acids, most of which are structurally and functionally unrelated to Ser. We hypothesize that these two Ser codons evolutionarily emerged independently, and the two Ser residues encoded by AGY and UCX were originally playing different roles in proteins. During evolution, the two different classes of Ser codons have been thoroughly mixed up by now so that, in most cases, one cannot distinguish clear differences in the use of these two Ser codons (4). Nevertheless, as discussed below, we can still find a disparity in the usage of these two Ser codons, because of either evolutional marks or distinct differences in the role of these two Ser codons at the level of translation or function. This has been shown previously in *E. coli*, where minor codons being used within the first 25 amino acids of a protein are involved in the regulation of gene expression (5, 6).

Based on these considerations, we analyzed a total of 4,225 protein-coding genes in *E. coli* (7). We found that the total number of AGY codons in the *E. coli* genome is 32,772, while the total number of UCX is 43,642, so that the ratio of AGY to UCX is 1 to 1.33 (http://www.kazusa.or.jp/codon). Since the

Significance

There are many questions related to the origin of life and how we came to rely on an almost universal system to encode all of life as we know it today. The genetic code is both robust and redundant, yet also full of interesting anomalies. Here we explore one of these anomalies, specifically the existence of two separate boxes/classes of serine codons, AGU/C and UCU/C/A/G. Unlike other synonymous codons encoding an amino acid, these codons for serine require a two-base substitution in order to go from one box to the other and remain a serine. Deciphering how this came to be will provide important insight into the origin of life and the genetic code.

Author contributions: M.I. and Y.I. designed research; R.T. and K.I. performed research; M.I., R.T., Y.I., and K.I. analyzed data; and M.I. and K.I. wrote the paper.

Reviewers: M.H.S., University of California San Diego; and D.S., Yale University.

The authors declare no competing interest.

See Profile on page 28543.

[1]To whom correspondence may be addressed. Email: inouye@cabm.rutgers.edu.

First published November 9, 2020.

Table 1. Codon table for all amino acids

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

Leu (blue), Arg (green), and Ser (orange), which are encoded by six codons, are highlighted in color. Note that UCX and AGY for Ser are located in two different positions so that their mutual exchanges require two base changes, UC ⟷ AG.

theoretical ratio of these two Ser codons is 1 to 2, AGY codons appear to be used disproportionally higher than UCX codons in the *E. coli* genome, with AGC being the most frequently used. In *E. coli*, of 1,000 amino acid residues in proteins, Ser appears 50.8 times (5.08%) (http://www.kazusa.or.jp/codon). Taking this as 100%, Ser encoded by AGC occupies 32.7%, followed by UCG (15.7%), UCA (15.3%), AGU (14.2%), UCU (11.2%), and UCC (10.8%), indicating that the usage of the AGC codon is distinctly higher than the other Ser codons. This seems to be consistent with the hypothesis that AGC is evolutionarily one of the most primitive codons for Ser (8). As discussed later, AGC was assumed to be derived from GGC for Gly, which originated from GGG for Gly. This hypothesis can be further extended to predict that all of the amino acids encoded by VGC or GVC (V = U, C, or A; at least one G residue in the first or the second position in the triplet), such as Arg (CGC), Ala (GCC), Val (GUC), and Asp (GAC), and also by VGG or GVG, such as UGG (a presumed termination codon at the early time of evolution); CGG for Arg, GUG for Val, GCG for Ala, and GAG for Glu are the most primitive amino acids during evolution in addition to Gly (GGC) and Ser (AGC). The other 13 amino acids (Leu, Ile, Met, Lys, Pro, Thr, Phe, Tyr, Trp, Cys, His, Asn, and Gln) together with Ser encoded by UCX are thus assumed to be secondarily evolved. We, therefore, hypothesize that the most primitive proteins were made of only the seven amino acids: Arg, Ala, Val, Asp, Glu, Ser (encoded by AGY), and Gly. Note that these seven presumed primordial amino acids include all of the fundamentally functional amino acids, such as a basic amino acid (Arg), acidic amino acids (Asp and Glu), a hydroxyl amino acid (Ser), two aliphatic amino acids (Ala, Val), and Gly. This hypothesis predicts that primitive enzymes, such as ribonucleases and proteases, consisted of only these seven primitive amino acids. This hypothesis conflicts with the proposal that 10 amino acids produced in prebiotic chemical experiments and also identified in meteorites; **Gly**, **Ala**, **Asp**, **Glu**, **Val**, **Ser**, Ile, Leu, Pro, and Thr (note that the amino acids shown in bold coincide with our proposed amino acids) are the primordial amino acids (3). We suggest that primordial primitive proteins did not necessarily have to contain all of the amino acids found in meteorites, such as Thr, Pro, Leu, and Ile.

In *E. coli*, Leu is the most abundant amino acid, as 10.17% of the total amino acids are occupied by Leu (http://www.kazusa.or.jp/codon). On the other hand, Ser occupies 5.08% of the total *E. coli* amino acids, so that the frequency of one-base random mutations on Leu codons is 2.00 (10.17/5.08) times higher than that of Ser codons. Notably, except for the U-to-C mutation at the third base of AGU, any other base changes (boldface) at all of the positions result in amino acid replacements, which could create eight codons for six different amino acids, **U**GU for Cys, **C**GU for Arg, **G**GU for Gly, A**U**U for Ile, A**C**U for Thr, AA**U** for Asn, and AG**A**(**G**) for Arg, representing several amino acids with differing properties (Cys, Arg, Ile, Asn, and Arg). In the case of Leu codons, CUG for Leu can be altered to four other codons by one-base replacement to remain a Leu codon (UUG, CUU, CUC, CUA), in addition to the following five amino acids: **A**UG for Met, **G**UG for Val, C**C**G for Pro, C**A**G for Gln, and C**G**G for Arg. It would seem that there is stronger selective pressure for serine to remain as AGY because a one-base substitution leads to more structurally and functionally different amino acids.

Based on the discussions above, the following hypotheses could emerge:

First, Poly-G is the most primitive polynucleotide, which is encoded to poly-Gly by being translated three bases at a time (8, 9).

Second, if a G residue in the poly-G mutates to a pyrimidine residue, U or C at the third (see boldface, below) position of a triplet Gly codon, the mutated codons still encode Gly. However, if to these mutated Gly codons, GG**U** or GG**C**, a second mutation occurs at the first position changing G to A (see boldface), the resulting triplet codons, **A**G**U** and **A**G**C**, encode Ser.

Third, UCU, UCC, UCA, and UCG for Ser codons evolutionarily emerged independently from AGU and AGC for Ser. We propose that **U**CU, **U**CC, **U**CA, and **U**CG codons (boldface indicates base changes) were secondarily derived from **G**CU, **G**CC, **G**CA, and **G**CG for Ala codons, respectively, which were derived from G**G**U, G**G**C, G**G**A, and G**G**G, respectively. Substitution of Ala residues with Ser not only makes a protein more hydrophilic but also, in some cases, may cause a protein to

acquire an enzymatic function or provide a site for protein modification, such as phosphorylation and acetylation.

Fourth, a point mutation on polyG that alters a G residue in polyG to U, C, or A creates nine new codons—UGG, GUG, GGU, CGG, GCG, GGC, AGG, GAG, GGA—which encode STOP, Val, Gly, Arg, Ala, Gly, Arg, Glu, and Gly, respectively, adding four new amino acids (Val, Arg, Ala, and Glu). In addition, a point mutation to codons, GGU(C) creates the following six new codons: UGU(C), GUU(C), CGU(C), GCU(C), AGU(C), and GAU(C), encoding Cys, Val, Arg, Ala, Ser, and Asp, adding three new amino acids, Cys, Ser, and Asp. Of these amino acids, Cys is considered to have emerged later in evolution (10), so that the following seven amino acids—Val, Ala, Gly, Asp, Glu, Ser, and Arg—are considered to be the most primitive amino acids, which would have made up the primordial proteins.

It is unlikely that the UCX Ser codons were derived from UGX because UGX represents two different amino acids as well as a nonsense codon: UGU/C for Cys, UGG for Trp, and UGA for a nonsense codon. As described above, UCX for Ser is assumed to be derived from GCX for Ala, which was derived from GGX for Gly, the most primitive amino acid. Therefore, it is further speculated that Ser residues encoded by AGU or AGC in proteins had originally different functions from Ser residues encoded by UCX. Since then, the two different sets of Ser codons have been thoroughly mixed up during evolution (4). However, as we discuss below, we are still able to detect the distinct different usages of the two Ser codons in some genes in *E. coli*.

Analysis of the Usage of Two Different Ser Codon Boxes, AGY and UCX, in Protein-Coding Genes in *E. coli*. The genetic codons for the Ser at the active center of proteases are encoded almost equally by AGY and UCX (4). These authors then argue that during evolution, extensive mixing-up has occurred between the primary Ser codons, AGY, and the secondary Ser codons, UCX, since the mutual exchanges between these Ser codons do not cause any effects on the functions of the proteins. However, since the exchanges between these two Ser codons, AGY and UCX, require at least two base changes, the exchanges must have happened through other amino acids, such as UCY (Ser) to ACY (Thr), and subsequently, ACY (Thr) to AGY (Ser), or reversely, AGY (Ser) to UGY (Cys) and UGY (Cys) to AGY (Ser) or AGY (Ser) to ACY (Thr) and ACY (Thr) to UCY (Ser). The exchanges between two Ser codons are unlikely to evolutionarily happen unless there were selective benefits associated with the mutations. Indeed, the active-site Ser residue is under an intense selective pressure to remain a Ser residue, which is the reason for the observation of the extensive mixing-up of the two Ser codons at the active center of proteases.

When all of the protein-coding genes in *E. coli* are analyzed, we find that there are a total of 37 out of 4,225 genes encoding proteins (7), consisting of more than 100 amino acid residues, in which Ser residues are encoded only by UCX. There ar a total of 23 genes encoding proteins, consisting of more than 100 amino acid residuees that use only AGY without using UCX. Furthermore, the sizes for those proteins having only UCX codons seem to be limited as there are only nine genes encoding proteins consisting of more than 200 amino acid residues. Interestingly, of 37 genes containing only UCX, there are several ribosomal proteins: Three 50S ribosomal proteins (L6 [4 UCX of a total of 177 coding triplets], L15 [7 of 144], and L18 [5 of 117]) and two 30S ribosomal proteins (S5 [7 of 167] and S11 [6 of 129]). There are 3 proteins having more than 10 UCX codons without AGY: Translation elongation factor Tu1 (10 of 394), glutamate/aspartate ABC transporter ATP binding subunit (12 of 241), and glyceraldehyde-3-phosphate dehydrogenase A (15 of 331). On the other hand, there are fewer ribosomal proteins in which Ser residues are encoded only by AGY: 50S ribosomal proteins L21 (3 of 103) and L16 (2 of 136). Notably, there are many genes containing more AGY than UCX, such as *rhsB* encoding a 1,411-residue protein containing 62 AGY and 28 UCX, and the 475-residue sensory histidine kinase, GlrK, containing 27 AGY and 14 UCX.

As mentioned above, the *gapA* gene for glyceraldehyde-3-phosphate dehydrogenase A, consisting of 331 amino acid residues, contains 15 Ser residues encoded only by UCX. If one considers that every Ser residue is encoded by chance by either AGY or UCX, the probability of having all 15 Ser codons encoded UCX would be $(4/6)^{15} = 0.00228$, suggesting that all 15 Ser codons in GapA were selected to be UCX during evolution. As discussed earlier, UCX Ser codons were likely derived from GCX codons for Ala by the G-to-U mutation, leading to another question as to why all of the 15 presumed Ala codons evolved to become Ser codons. A possible explanation may be that the GapA protein initially existed in a hydrophobic environment, which might have gradually become more hydrophilic, forcing the mutation of the Ala residues to Ser residues. On the other hand, when we analyzed the genomes of nine additional bacteria and three archaea for their usages of the two Ser codons, it was found that AGY and UGX usage was generally consistent with that of *E.coli.* Analysis of five different genes (*rpmH*, *adk*, *gapA*, *rplU*, and *ribE*) that used only UCX or AGY in *E.coli* in these 12 organisms revealed that there was extensive mixing-up of the two UCX and AGY Ser codons within these genes, which likely happened during evolution (4). As for tRNAser, there are five independent genes on the *E. coli* genome; however, all of them show very high homology, indicating that they are evolutionarily related, sharing the same origin (12).

Recently, it was shown that the amino acid residues in human proteins that are conserved in around 100 vertebrates have a greater number of substitutions involving Ser encoded by UCX, Pro (CCX), and Ala (GCX). On the other hand, the less-conserved residues tend to be filled with Ser encoded by AGY, Glyc (GGX), and Asn (AAY) (11). These less-conserved residues indicate that there were selective pressures to choose one of the sets of Ser codons according to the evolutional history of functions and structures of proteins.

Concluding Remarks

Looking at the codon table (Table 1), we seem to be able to decipher hidden stories about how genetic codons evolved. Based on the hypothesis that the simplest and thus the most primitive amino acid among the 20 amino acids is GGX or Gly, the codons for other amino acids are proposed to have evolved from GGX. In the second step of codon evolution, new sets of the codons for seven amino acids emerged: Val (GUX), Ala (GCX), Asp (GAY), Glu (GAZ), Ser (AGY), Arg (AGZ and CGX), and Gly (GGX). These were derived from the Gly codons by a single-base substitution either at the first or the second base position. In the third step, the codons for the remaining 13 amino acids evolved from the codons for the primitive seven amino acids. However, the second set of codons for Ser (UCX) is an exception since the other set of codons for Ser (AGY) already existed. We propose that the second set of Ser codons (UCX) independently evolved from the codons for Ala (GCX), functionally altering the hydrophobic nature of Ala residues in proteins to hydrophilic. Therefore, Ser residues in proteins encoded by UCX were originally different from Ser residues encoded by AGY. Further examination of other bacteria and also archaea for the usages of the two different Ser codon boxes will likely further shed light on the question on their evolutional origins.

Data Availability. All study data are included in the article.

ACKNOWLEDGMENTS. The authors are grateful to Drs. M. Hampsey and M. Travisano for their critical reading of the manuscript.

1. M. Nirenberg, Protein synthesis and the RNA code. *Harvey Lect.* **59**, 155–185 (1965).
2. H. G. Khorana, Polynucleotide synthesis and the genetic code. *Fed. Proc.* **24**, 1473–1487 (1965).
3. E. V. Koonin, A. S. Novozhilov, Origin and evolution of the universal genetic code. *Annu. Rev. Genet.* **51**, 45–62 (2017).
4. I. B. Rogozin *et al*., Evolutionary switches between two serine codon sets are driven by selection. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 13109–13113 (2016).
5. G. F. Chen, M. Inouye, Suppression of the negative effect of minor arginine codons on gene expression; preferential usage of minor codons within the first 25 codons of the Escherichia coli genes. *Nucleic Acids Res.* **18**, 1465–1473 (1990).
6. G. T. Chen, M. Inouye, Role of the AGA/AGG codons, the rarest codons in global gene expression in Escherichia coli. *Genes Dev.* **8**, 2641–2652 (1994).
7. P. Hu *et al*., Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. *PLoS Biol.* **7**, e96 (2009).
8. H. S. Bernhardt, W. M. Patrick, Genetic code evolution started with the incorporation of glycine, followed by other small hydrophilic amino acids. *J. Mol. Evol.* **78**, 307–309 (2014).
9. H. S. Bernhardt, W. P. Tate, Evidence from glycine transfer RNA of a frozen accident at the dawn of the genetic code. *Biol. Direct* **3**, 53 (2008).
10. J. T. Wong, A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci. U.S.A.* **72**, 1909–1912 (1975).
11. G. W. Schwartz, T. Shauli, M. Linial, U. Hershberg, Serine substitutions are linked to codon usage and differ for variable and conserved protein regions. *Sci. Rep.* **9**, 17238 (2019).
12. K. L. Roy, D. Soll, Purification of five serine transfer ribonucleic acid species from *Escherichia coli* and their acylation by homologous and heterologous seryl transfer ribonucleic acid synthetases. *J Biol Chem.* **245**, 1394–1400 (1970).