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

Masayori Inouyea,1, Risa Takinoa, Yojiro Ishidaa, and Keiko Inouyea

aDepartment 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.

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. Particularly, 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 GCX 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.

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.

Published under the PNAS license.

1To whom correspondence may be addressed. Email: inouye@cabm.rutgers.edu.
appear to be used disproportionately 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 occurs 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 GCC 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, GCC for Ala, and GAG for Glu are the most primitive amino acids during evolution in addition to Gly (GCC) 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, UGU for Cys, CGU for Arg, GGU for Gly, AUU for Ile, ACU for Thr, AAU for Asn, and AGAA(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: AUG for Met, GUG for Val, CCG for Pro, CAG for Gln, and CGG 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 these mutated Gly codons, GGU or GGC, a second mutation occurs at the first position changing G to A (see boldface), the resulting triplet codons, AGU and AGC, encode Ser.

Third, UCU, UCC, UCA, and UCG for Ser codons evolutionarily emerged independently from AGU and AGC for Ser. We propose that UCU, UCC, UCA, and UCG codons (boldface indicates base changes) were secondarily derived from GCC, GCA, and GCC for Ala codons, respectively, which were derived from GGU, GCC, GGA, and GGG, 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, UGU, GGU, CGG, GCC, GGC, AGG, GAG, GGA—which encode STOP, Val, Gly, Arg, Ala, Gly, Arg, Glu, and Gln, 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: UGCU(C), GUUCU(C), CGCU(C), GCCU(C), AGUCU(C), and GAUCU(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 GGY for Gly, the most primitive amino acid. Therefore, it is further speculated that Ser residues encoded by AGY 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 differences of the UCX and AGY 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 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, occur 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 UCY (Cys) and UGY (Cys) to AGY (Ser) or AGY (Ser) to ACY (Thr) and ACY (Thr) to UCY (Ser). The exchanges between these 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 are a total of 23 genes encoding proteins, consisting of more than 100 amino acid residues 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 (L5 [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 TuI (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 nsbH 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 one of two Ser codons (AGY or UCX), the probability of having all 15 Ser codons encoded UCX would be (4/6)^15 = 0.000228, 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 a G-to-U mutation, leading to the question as to why all of the 15 presumed Ala codons evolved to become Ser codons. A possible explanation is 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 (adk, gntH, dhaA, gapA, and rplU) 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 evolutionary 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 GGY or Gly, the codons for other amino acids are proposed to have evolved from GGY. In the second step of codon evolution, new sets of the codons for seven amino acids emerged: Val (GUX), Ala (GCX), Asp (GAX), 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 evolutionary 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.