Fitting the standard genetic code into its triplet table

Michael Yarus[a,1]

[a]Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, CO 80309-0347

Edited by Harry F. Noller, University of California, Santa Cruz, CA, and approved July 16, 2021 (received for review October 8, 2020)

Minimally evolved codes are constructed here; these have randomly chosen standard genetic code (SGC) triplets, completed with completely random triplet assignments. Such "genetic codes" have not evolved, but retain SGC qualities. Retained qualities are basic, part of the underpinning of coding. For example, the sensitivity of coding to arbitrary assignments, which must be < ∼10%, is intrinsic. Such sensitivity comes from the elementary combinatorial properties of coding and constrains any SGC evolution hypothesis. Similarly, assignment of last-evolved functions is difficult because of late kinetic phenomena, likely common across codes. Census of minimally evolved code assignments shows that shape and size of wobble domains controls the code's fit into a coding table, strongly shifting accuracy of codon assignments. Access to the SGC therefore requires a plausible pathway to limited randomness, avoiding difficult completion while fitting a highly ordered, degenerate code into a preset three-dimensional space. Three-dimensional late Crick wobble in a genetic code assembled by lateral transfer between early partial codes satisfies these varied, simultaneous requirements. By allowing parallel evolution of SGC domains, this origin can yield shortened evolution to SGC-level order and allow the code to arise in smaller populations. It effectively yields full codes. Less obviously, it unifies previously studied chemical, biochemical, and wobble order in amino acid assignment, including a stereochemical minority of triplet–amino acid associations. Finally, fusion of intermediates into the final SGC is credible, mirroring broadly accepted later cellular evolution.

RNA world | translation | codon | anticodon | evolution

The form of the standard genetic code (SGC) offers authoritative information about its origin. By calculating evolved coding tables via different pathways (1), the SGC's implications for its creation can be investigated. More frequent SGC-like results quantitatively signal superior explanations. Consequently, initial hypotheses about code descent can be improved. In fact, respecting Bayes' theorem, multiple successful explanations rapidly strengthen an accurate hypothesis by Bayesian convergence (2).

The existing result is late Crick wobble (3). "Late" implies that NNR (R = purine) and NNY (Y = pyrimidine) wobble was deferred, being preceded by unique triplet pairing assignments. Unique triplet pairing does not require support from a highly evolved allosteric ribosome (4, 5); it does not require a specific, highly optimized tRNA anticodon loop-and-stem structure (6, 7) or control of varied isomerization of wobble-paired bases (8). Accurate Crick wobble (9), with these multiple requirements, would therefore likely be a later, more modern code refinement. Late wobble also shows superior ability to fill an SGC-like coding table and offers more probable SGC access to evolving codes (1, 3). Moreover, the first wobble (the SGC necessarily uses ambiguous wobble coding) probably resembled simplified Crick wobble, defined here as translation of NNY and NNR codons, each with one adaptor RNA (9). For comparison, superwobble, translation of four NNY/R codons with one unmodified adaptor (10), is less probable because it less frequently yields the SGC (3). Other evolutionary routes to SGC-like coding can be evaluated comparably, within this simulated framework.

Results

Random, Minimally Evolved Codes. Minimally evolved codes are random coding tables or they differ in only defined ways from randomly assigned coding tables (*Methods*). They have determined fractions of randomly chosen SGC codons, completed with completely random assignments. A property appearing in such a minimally evolved state is likely intrinsic to code evolution. This construction therefore defines essential evolutionary problems for the SGC and allows evaluation of claimed solutions.

Inappropriate Assignments. The simplest among these inevitable requisites arises from overall sensitivity of code evolution to nonspecific triplet assignment. Fig. 1*A* plots assignment accuracy; the probability of codes with 0, ≤1, ≤2, ≤3, or ≤4 misassignments (abbreviated "mis") relative to the SGC. Code accuracies are shown versus the fraction (probability of random assignment, Prand) of random triplet meanings (20 amino acids/termination/initiation) rather than canonical SGC triplet assignments. Fig. 1*A*'s points average 10^4 independent coding table evolutions, and thus are not tied to any particular choice of SGC or randomized codons. Accuracy is determined after encoding 20 functions, then late Crick wobble (1, 3, 11). Fig. 1 results resemble previous calculations of code order (spacing, distance, and chemical order taken together) versus Prand. In particular, approaching SGC order or acquiring SGC-like assignments (Fig. 1*A*) requires that random codon assignment be < ∼15% of the total, preferably < ∼10% (1). SGC-like coding is very sensitive to Prand, and sensitivity increases as resemblance to the SGC increases.

Inappropriate Assignments: A More Revealing Graphic. These conclusions are fortified by a more informative plot. Fig. 1*B* posits a minimally evolved coding table using random assignment probabilities (Prand). Predicted code accuracies are approximated in Fig. 1*B* using the binomial distribution for a coding table with the same triplet occupancy in Fig. 1*A* (*Methods*).

Fig. 1*C* includes data for minimally evolved codes as in Fig. 1*A*, but without mutational capture of initial assignments by codons related by single mutational changes. Clearly these minimally evolved codes, which have late Crick wobble, and all other characteristics of Fig. 1*A* except capture, greatly resemble the true

Significance

The standard genetic code (SGC) is common to all sufficiently explored Earth biota, suggesting a common origin for protein biosynthesis in every known organism. The ancient events leading to this near-universal biological characteristic are manifestly of scientific importance. The accompanying text presents a detailed pathway for formation of the SGC, as well as a method for identifying essential origin events. Significantly, the SGC can evolve using only well-characterized chemical, biochemical, and physical mechanisms, paralleling other known evolutionary transitions.

Author contributions: M.Y. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The author declares no competing interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]Email: Michael.Yarus@Colorado.edu.

Published August 30, 2021.

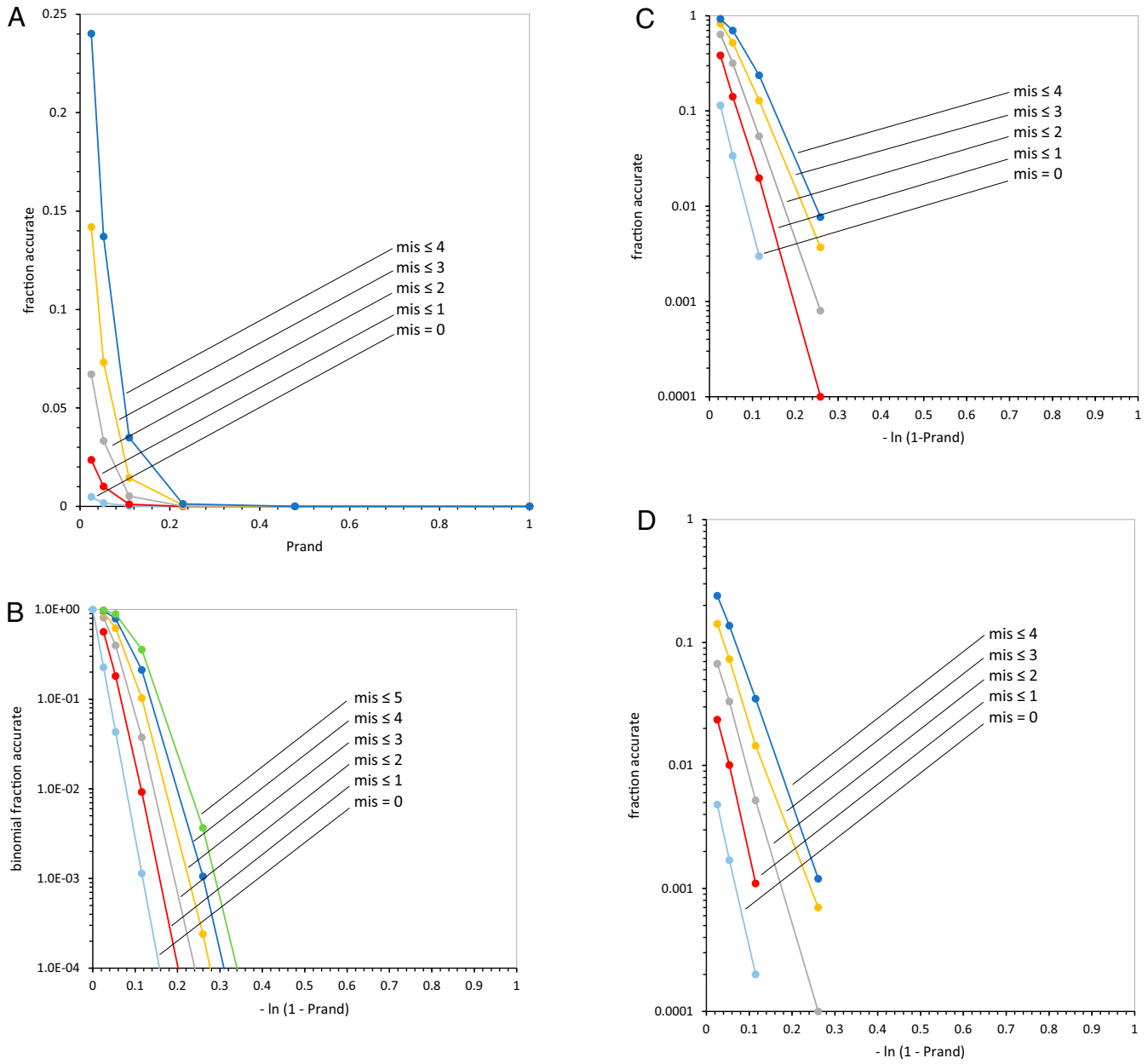

Fig. 1. (*A*) Sensitivity of code evolution to randomness. Fraction accurate = the fraction of 10^4 evolutions that have indicated accuracy. Accuracy is measured by counting misassignments relative to the SGC; mis = 0 (no errors), mis ≤ 1 (single error), mis ≤ 2, mis ≤ 3, and mis ≤ 4. Prand = fraction completely random assignment; thus (1 – Prand) = fraction SGC assignments, chosen randomly from an SGC table. Results are for late Crick wobble evolution to 20 encoded functions. (*B*) A better portrayal: calculated binomial sensitivity to randomness. Calculated binomially distributed probabilities of choosing 58.6 triplets (the mean in *A*) with varied Prand (*Methods*). Accuracy color coding appears as in *A*. (*C*) A better portrayal: a minimally evolved code and randomness: late Crick wobble, no mutational capture. Evolution of codes as in *A* (with late Crick wobble), but without mutational capture of assignments. Accuracy color coding appers as in *A*. (*D*) A better portrayal: sensitivity of late Crick wobble evolution to randomness. Data for complete evolution of late Crick wobble as in *A*, but plotted as in *B*. Accuracy color coding appears as in *A*.

randomness of Fig. 1*B*. However, Fig. 1*C* makes it much clearer that the rarity of accurate codes is innate. As random assignment increases [Prand as *x* axis trends similarly to −ln (1 – Prand); *Methods*], the frequencies of accurate codes, particularly those identical to the SGC (mis = 0), fall exponentially and indefinitely.

Fig. 1*D* replots data from Fig. 1*A* for complete code evolution, now including mutational capture. Comparison to Fig. 1*C* shows that realistic evolution is yet more sensitive to randomness, yielding ~10-fold fewer accurately assigned codes (Fig. 1*D*). Fig. 1 *C* and *D* taken together show that this increased sensitivity is attributable to mutational capture in Fig. 1*D*. This is a first indication of the negative effects of difficult fit among captured wobbling triplets, explained below.

Completion Complications: Kinetics. Codon assignment necessarily slows near code completion, because assignment decay (which decreases assigned triplets) is speeding up, and mutational capture and triplet assignment (which increase assigned triplets) are slowing down (1). Slowed late assignments suggest why definitive initiation and termination mechanisms were assigned late, by an

independent selection, as suggested by their mechanistic differences between life's domains (11).

Fig. 2 shows that slowed code completion is intrinsic, as expected from its origin in assignment kinetics. Fig. 2*A* is the behavior of a minimally evolved code as it approaches complete assignment (17 to 22 assigned functions). The figure, showing time in passages (1), plots requirements for each level of assigned coding function. And, as an indicator of the efficiency of the assignment process, the mean number of assignment decays for a triplet to acquire its final meaning (decays/assignment) is also shown. Fig. 2*A*'s minimally evolved coding tables are filled randomly, but with no wobble or mutational capture. Such coding still slows near completion, with 36-fold as much time required for 22 functions as for 20, requiring 53-fold as many decays per assignment for 22 encoded. For such random filling with 10% random assignments, triplets must be multiply assigned (3.5 times on average) to reach complete 22-function coding.

Code completion becomes much more burdensome with continuous Crick wobble, even without mutational capture, as in Fig. 2*B*. Wobble from the start of code evolution increases both 20 to 22 function time and complexity by about sevenfold. In Fig. 2*B* the average triplet assignment has decayed 24.2 times in order to encode 22 functions.

If mutational capture of neighboring triplets for related assignments is added to Fig. 2*B*, as in Fig. 2*C*, then code completion via continuous Crick wobble is yet more hindered. To encode 22 functions, assigned triplets have decayed an average of 234 times, more than 100-fold exceeding that at 20 functions. And, time to evolve to 22 from 20 functions is >100-fold the time to 20 encoded. Assignment of these latter triplet meanings would be tortuous, each decaying many times before 22 functions are attained. These results reproduce and extend previous comparisons in which early wobble, and early wobble to an extended range of triplets, resulted in delay and inefficient evolution (1, 3). Increased effort for completion—from near-random filling (Fig. 2*A*) to continuous wobbling alone (Fig. 2*B*) and increased further for continuous wobbling in captured codons (Fig. 2*C*)—further exemplifies increasing difficulty in fitting coding assignments.

Completion Complications: Fitting. However, fitting difficulties continue even if encoding 22 functions is avoided by selecting the last two functions later. In the original discussion of completion complications (1), difficulties are said to be both kinetic and due to a large universe of possible codes. Now we turn to the second kind of completion complexity, describing assignment fit during the first 20 encodings.

Difficult Fitting: Encodings that Cannot Overlap. Fitting implies difficulty placing code domains, like wobbling sets of codons that must preserve unambiguous meanings, into a finite, fixed coding space. Such encodings are of differing size and complexity. In this work (Fig. 3*A*), triplet assignments can be unique (red),

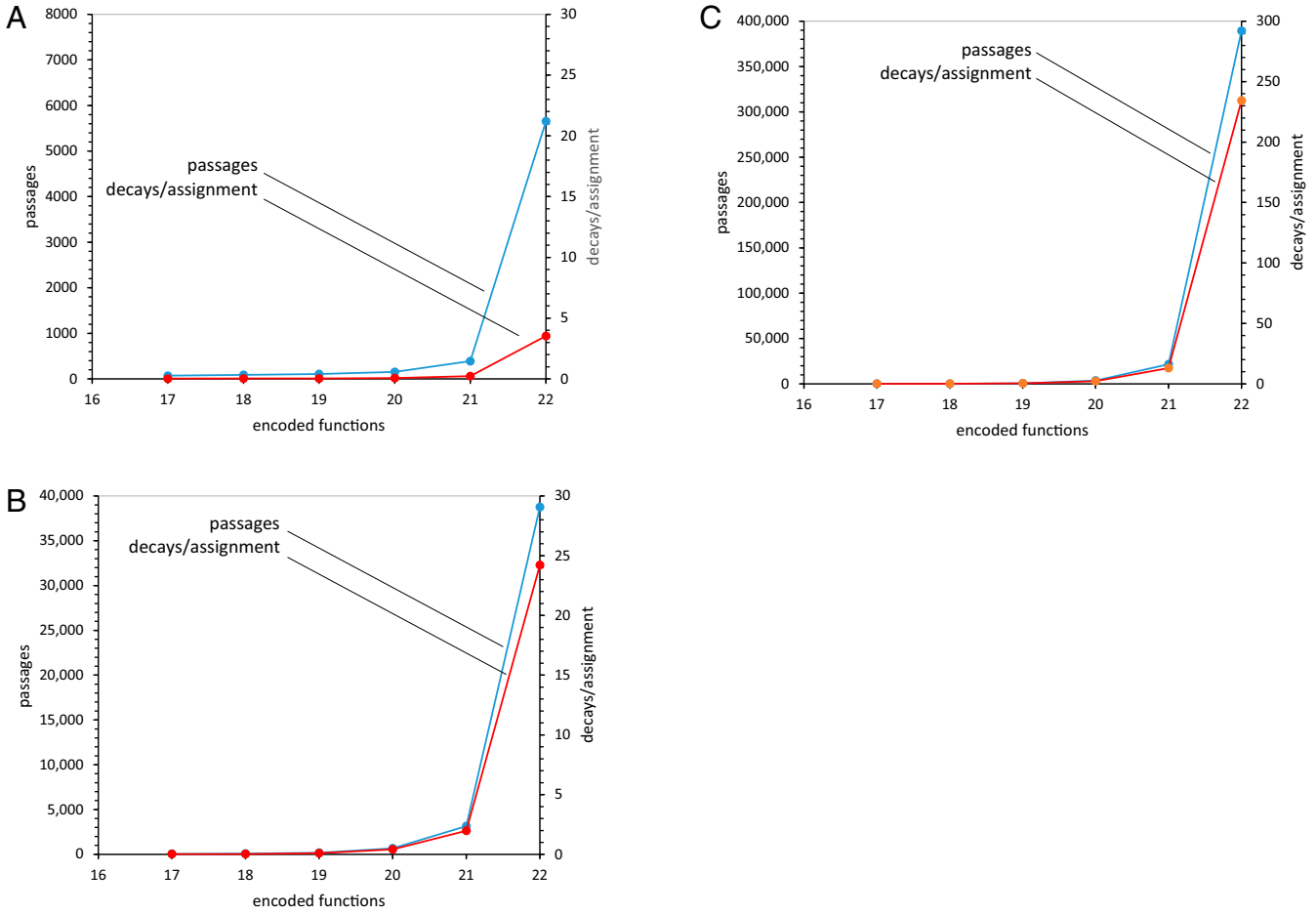

Fig. 2. (*A*) Completion complications: within a minimally evolved code. Means for 10^3 coding tables constructed with Prand = 0.1 and no wobble, no mutational capture. Passages are computer visits to an evolving coding table, proportional to time (1). (*B*) Completion complications: with wobble. Means for 10^4 coding tables evolved with Prand = 0.1 and continuous Crick wobble (that is, throughout evolution), no mutational capture. (*C*) Completion complications: with wobble and mutational capture. Means for 10^3 coding tables evolved with Prand = 0.1 and continuous Crick wobble, with mutational capture.

Crick wobbling (yellow, in an example initiated at AAU), superwobbling (orange, in an example initiated at GAC), or the size of the complete mutational neighborhood for capture by single mutation (blue, for Crick wobbling initiated at UUG as an illustration: Fig. 3*A*).

Pmut is the probability, per passage per neighboring triplet pair, that a chosen assigned codon will confer its meaning or a related meaning on an unassigned triplet one mutation distant (*Methods*). As Pmut increases for late Crick wobble in Fig. 3*B*, related assignments can increasingly spread across areas like Fig. 3*A*'s blue areas, which define a mutational neighborhood for Crick wobble initiated at the italicized *UUG* codon. Thus, as Pmut varies 40-fold in Fig. 3*B*, such mutational capture/related assignment goes from rare to major evolutionary event, the latter being usual in these calculations (1).

Fig. 3*B* shows that fitting is a substantial quantitative consideration. It plots probability of accurate assignment. Mean levels of codon misassignment relative to the SGC are plotted for 10^5

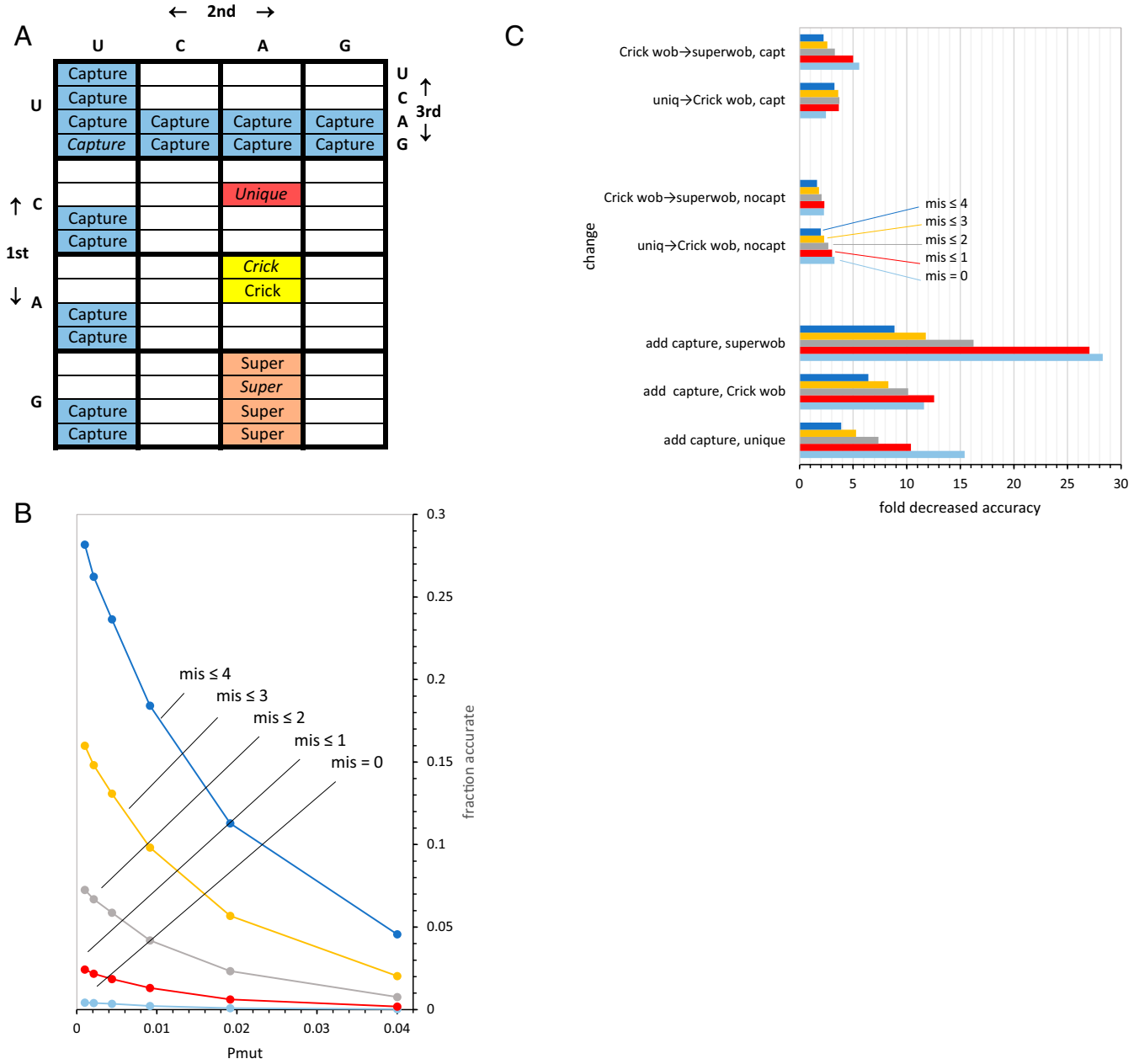

Fig. 3. (*A*) Mutational capture domains: examples in coding table context. Unique coding assigns only the target triplet. Crick wobble assigns two NNR or NNY triplets, where R = purine and Y = pyrimidine. Super = superwobble (10), which assigns four triplets, NNR/Y. Capture = all codons related to example codon UUG by single mutation, followed by Crick wobble. (*B*) Effect of mutational capture with Crick wobble on assignment accuracy. Pmut = probability, per passage, per eligible triplet pair, for mutational capture with Crick wobble; data are means of 10^5 evolutions. Pmut varied from 0.001 to 0.04. Other probabilities are as in *Methods*. Accuracy color coding appears as in Fig. 1*A*. (*C*) Fold accuracy decrease due to expansion of wobble capture domain: ± capture, with unique assignment, Crick wobble, superwobble. Mean effect, in 10^5 evolutions, of the "change" described in titles on the *Left*. Change description titles indicate, first, the conditions being compared, in a minimally evolved background, then other relevant but constant conditions are cited. For example, "uniq→Crick wob, capt" plots the change in accuracy observed when Crick wobble is added to a minimally evolved code that used unique assignment, both populations using mutational capture. Accuracy color coding appears as in Fig. 1*A*.

evolved codes, varying as related assignment for neighborhood codons varies. Data are collected after assignment of 20 functions, for reasons described above, and because 20 functions are acquired near the time when best near-SGC order occurs, and therefore near the time when the SGC itself was probably selected (11).

Negative Effect of Assignment to a Neighborhood. Mutational capture always decreases the mean accuracy of coding. Codes with four or fewer differences from SGC coding decrease >6-fold with mutational capture increases of 40-fold (Fig. 2*B*). The negative capture effect becomes more important for greater accuracy; codes that completely emulate SGC assignment (mis = 0) are 14-fold less frequent for the same increased capture.

Fig. 3*C* generalizes these findings to the complete range of fittings depicted in Fig. 3*A*. Mean fold decreases in accuracy, measured by changed misassignment levels in 10^5 evolutions to 20 encoded functions, are shown for the introduction of Crick wobble and superwobble into uniquely assigned codes, with (capture with the usual Pmut = 0.04, *Methods*) and without (Pmut = 0.001) assignment of related encodings to mutationally related codons (Fig. 3*A*).

Decrease in Accuracy Increases when More Precision Is Demanded. Bundles of data in Fig. 3*C* present similar data for accuracy, from four misassignments (dark blue) at the top of each bundle of columns, to complete SGC verisimilitude (0 misassignments, light blue) at bundle bottom. A glance at the figure reveals greater decreases in accuracy as accuracy itself increases. Decreases are larger at mis = 0 than mis ≤ 4. Thus the trend in Fig. 3*B* is more general; fitting becomes more difficult as SGC resemblance increases or misassignment decreases.

Decrease in Accuracy Increases with Size of the Domain Placed. If unique assignment is changed to Crick wobble, or similarly, Crick wobble is changed to superwobble, the size of potential wobble domains doubles (Fig. 3*A*); one to two triplets and two to four triplets, respectively. At the *Top* of Fig. 3*C*, these similar wobble expansions have 2- to 5-fold effects on misassignment, whether their expanded wobbles are superposed on a system without mutational capture (*Middle* data bundles) or a system with mutational capture (*Top* bundles). On the other hand, in codes evolving with capture, which distribute their wobbles over a larger domain (*Top*), accuracy effects are larger than when a mutational capture domain (blue, Fig. 3*A*) is not engaged (*Middle* bundles, Fig. 3*C*). Disruptive effects ranging up to 27-fold occur for introduction of large capture domains (blue, Fig. 3*A*). Negative effects on assignment accuracy in these large domains clearly increase with wobble domain size: capture of unique codons is less destructive of accuracy than capture with Crick wobble; Crick wobble is less obstructive than capture with superwobble (*Lower* three bundles, Fig. 3*C*).

Summary of Difficult Fitting. Accuracy penalties for fitting wobble domains intensify as domains grow and also as more accuracy is demanded. Both effects make intuitive sense and also agree with and generalize earlier results. Previously, among populations of evolving late Crick wobble codes, the minority of codes that most resembled the SGC had also made fewer mutational captures (1). Specific comparisons of unique and Crick wobble (1) and Crick and superwobble (3) previously favored the smaller, simpler forms. Moreover, these results (Fig. 3 *C*, *Top*) recapitulate the previous numerical superiority of Crick over superwobble (3)—prior results now recognizable as one sign of a more general fitting effect (Fig. 3*C*).

What evolutionary routes minimize Fig. 3's negative fitting effects, making SGC-like assignments more accessible? Framing this as a "fitting problem" immediately suggests a solution: fragments of complementary shape might fit smoothly. In Fig. 4, several well-known code substructures unexpectedly exhibit this unifying fitting property. Accordingly, primordial ordered code fragments can join softly to minimize, or even entirely eliminate wobble impacts on fit.

Primordial Ordered Code Fragments: A Row of Early Amino Acids. Eigen et al. (12) noted that the most prominent amino acids from sparked gases designed to emulate primitive reducing atmospheres (13) also have a unique coding position. These amino acids: Val, Ala, Asp, Glu, and Gly, are the present occupants of the GNN row of the SGC. These findings therefore lend themselves to theories, including Eigen's, that a primordial code would have encoded these chemically "primitive" amino acids within the corresponding row of the coding table, using first codon position G somewhat as shown in Fig. 4*A*. Extensive further work on chemical properties, such as the free energy cost of synthesis in seawater (14), strikingly confirms that these five amino acids could be prevalent before biosynthesis. This grouping was also strengthened by Taylor and Coates (15), who noted that the same amino acids and their codons could also be classed as early, or sometimes, precisely the earliest amino acids produced from major glycolytic and citric acid cycle intermediates (as for Ala, Val, Asp, and Glu). Thus chemical and biosynthetic indications concur that Val, Ala, Asp, Glu, and Gly may have been early assignments to their present GNN code row (16).

Primordial Ordered Code Fragments: Synthesis and Rows. A correlation between synthesis and SGC rows can be extended from the above five amino acids, arguably abundant on the early Earth, to at least 16 of 20 amino acids ultimately encoded by the SGC. Fig. 4*B* shows that biosynthetic origin and row coding are related for presumably later-arising amino acids, which required evolution of a biosynthetic pathway. First, in Fig. 4*B*, cell colors indicate likely origins from a basic metabolic intermediate (15): green for derivation from glycolytic phosphoglycerate, blue from glycolytic phosphoenolpyruvate, yellow from citric acid cycle oxaloacetate, and pink from citric acid cycle α-ketoglutarate. Different anabolic origins clearly tend to segregate into SGC rows, though a row relation is not completely observed. A more comprehensive summary would include a minority of precursor–product relations that are related by first codon position (column) mutation (Fig. 4*B*). Nevertheless, frequent ordering by row suggests that the code was formed during the period when biosynthesis itself was also being established. In addition, biosynthesis approximately conserved the tendency initiated by early availability: amino acids encoded in one chemical or biochemical era tend to find assignments within an SGC row.

Moreover, in Fig. 4*B*, text shading of amino acid names roughly indicates the order of synthesis of the amino acids: white → gray → black. Thus white Asp (from oxaloacetate) is the precursor of gray Thr, which in turn is a precursor to black Ile (15, 17). Pooling such relations, Fig. 4*B* spans examples in columns and also rows. Fig. 4*B*'s rows and columns taken together thereby exhibit the basis for the coevolution theory of SGC origins, in which the extension of early biosynthesis results in the assignment of closely related triplets (requiring only single mutations) to encode newly synthesized amino acids (17–19). Notably, such triplet concessions are not limited to rows, but have occurred in the first codon nucleotide (e.g., Arg), the second nucleotide (e.g., Tyr, from Phe), and the third nucleotide (many examples, e.g., Ser, Pro, and Thr).

Likely ancient chemistry and biosynthetic pathways can be fit together smoothly (20): Di Giulio noted that the coevolution of assignments and amino acid synthesis (17) can grow from initial assignment based on ancient availability of the GNN row: Val, Ala, Asp, Glu, and Gly (Fig. 4*B*).

Primordial Ordered Code Fragments: Columns and Amino Acid Chemistry. Woese et al. (21) suggested that the code might have originated as columns, particularly with hydrophobic amino acids partitioned into

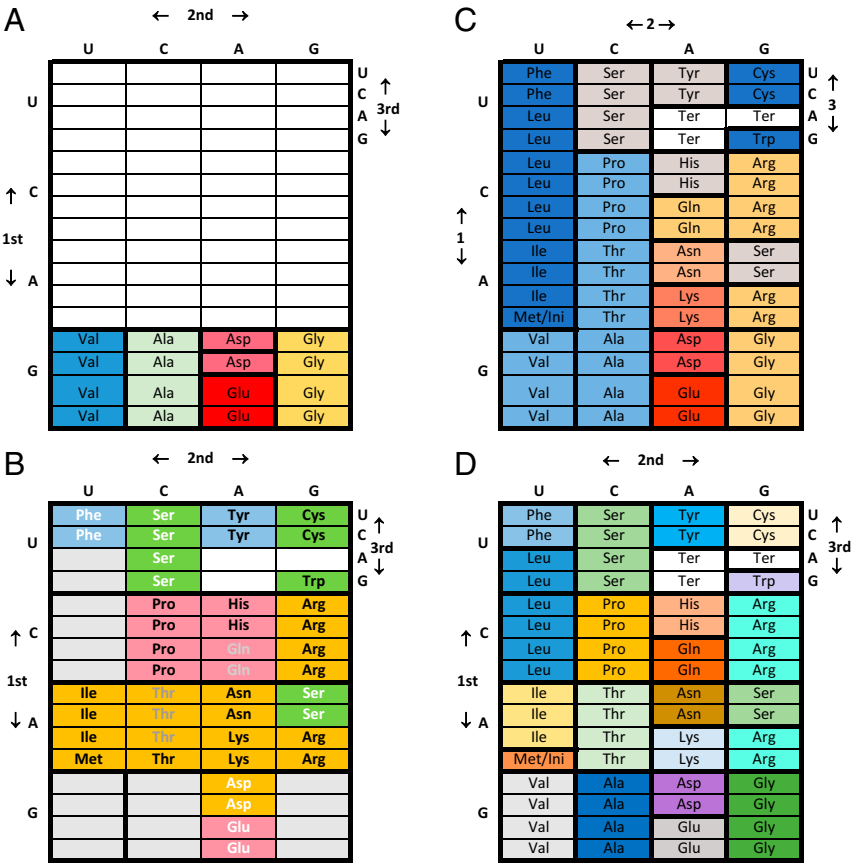

Fig. 4. (*A*) Elements of ordered coding: natural amino acids. Coding table for amino acids arguably available from ancient chemical sources, colors are arbitrary. (*B*) Elements of ordered coding: biosynthetic groups. Coding table for amino acids derived from glycolytic phosphoglycerate, green; from glycolytic phophoenolpyruvate, blue; from citric acid cycle oxaloacetic acid, yellow; and from citric acid cycle α-ketoglutarate, pink (15). Text colors indicate order within an amino acid biosynthetic pathway: white, first; gray, second; black, last. (*C*) Elements of ordered coding: chemical groups. Coding table colored for hydrophobic chemical character, grouping amino acids within one polar requirement unit (21). From hydrophobic to hydrophilic: dark blue, light blue, gray, yellow, orange-yellow, red, and deep red. (*D*) Elements of ordered coding: wobble groups. Coding table is colored similarly for identical amino acids whose triplets are linked by wobble, that is, by third codon nucleotide variation. Individual colors are arbitrary.

SGC columns with pyrimidines in the second codon position. Analysis of amino acid chemistry in the four coding columns (16) agrees, though the third (NAN) and fourth (NGN) columns of the SGC are less readily rationalized than the first (NUN) and second NCN) columns. An ancient triplet code in which the second position is the initially meaningful one is suggested by ref. 22; the result is a "2-1-3" model in which the SGC develops column by column, starting with the second column, then first, then third column. These ideas are approximated in Fig. 4*C*, which shows the SGC displayed by chemical character, using polar requirement (23, 24) as the display index. Each color represents one unit in PR, with blue the most hydrophobic (PR = 4.01 to 5.0), then light blue, gray, yellow, orange, and red the most hydrophilic amino acid (PR = 13.01 to 14.0). The SGC is plainly composed of large areas with very similar PR, amino acids whose PR is within one unit. There is a strong tendency to columns that differ, but that tend to have internally similar chemical character. A column's tendency is not always to equality. While the third column varies continuously in PR, its structure arguably follows a continuous gradient of amino acid chemical character, top hydrophobic to bottom hydrophilic (11).

Primordial Ordered Code Fragments: Wobble Domains. It would be overlooking a major kind of SGC order to neglect the abundantly ordered wobble position, however familiar it is (Fig. 4*D*). Except for the assignments of Met/Ini and Trp, wobble ordering is universal in the SGC. The most likely explanation of SGC evolution (the Bayesian convergence) (2) would simultaneously account for all coding regularities. Chemical character (PR) should often be conserved in columns (Fig. 4*C*), chemical (Fig. 4*A*) and biochemical origins together should tend to follow rows (Fig. 4*B*), and wobble behavior should almost invariably capture third nucleotide domains (Fig. 4*D*).

Discussion

Nonrandomness and Stereochemical Effects. Initial specific codon assignments are often treated as synonymous with assignments based on a definite chemical relation, like binding between RNAs containing coding triplet sequences and amino acids (25). Such specific relations are called "stereochemical." In Fig. 1, overall randomness in underlying assignments must be limited to < ~10%, in order to observe codes with 20 assignments that are also SGC-like in their encoding (1). Sensitivity to randomness (Fig. 1) grows from inevitable combinatorics (1). Any highly ordered code, realized via any evolutionary pathway whatever, has reached an improbable destination (1). Rare SGC-like outcomes require explicit justification.

Stereochemical Origins for an SGC. On one hand, side chain- and stereo-specific RNA binding sites for amino acids have long been known (26), natural RNA–amino acid binding sites exist (26, 27), bioinformatic evidence for amino acid–codon relations is plentiful (28), and a group of amino acid–cognate RNA sites has

been experimentally selected (2, 25, 29). But while cognate coding triplets recur in amino acid binding sites unexpectedly frequently, they are still sparse. Selection experiments on eight amino acids potentially encoded by 25 codons/25 anticodons find nine cases in which the cognate coding triplet nucleotides are improbably frequent in RNA binding sites, and are essential functional elements for amino acid affinity (25). Such coding triplets are especially prominent in the simplest, and therefore likely the most primitive, specific binding sites (25, 30). In total, 7 anticodons and 2 codons have been detected in side chain–specific oligoribonucleotide binding sites. Thus, 28% of tested cognate anticodons and 8% of tested cognate codons appear as essential functional sequences in RNA binding sites for their cognate amino acid.

Such sparseness is unsurprising: detailed molecular interactions required for an indispensable code triplet role within specific RNA-bound amino acid domains would not be expected for any but a minority of tertiary structures. So, a stereochemical code origin must explain: even given multiple experimental cases of functional coding triplets, how did > ~90% of 61 triplet–amino acid assignments become ordered?

A Credible Stereochemical Solution. The SGC can assemble from multiple smaller domains, each with nonrandom structures deriving from one, or a few, founding stereochemical interactions. A plausible path to the SGC then requires a way to merge these elements.

A Different SGC Presentation. A more realistic visualization of codon–amino acid relationships is useful. The familiar flat rectangle compresses a three-dimensional relation (for three triplet nucleotides) to two dimensions. Explicitly recognizing the third dimension (Fig. 5) appreciably changes the code's appearance, and the added dimension clarifies a complex origin (1). In Fig. 5, the first two codon nucleotides extend, U to C to A to G, frontward (first nt) and left to right (second nt). This creates sheets with the wobble nucleotide (third nt) varying in UCAG order vertically, spanning stacked first/second nucleotide planes.

Color selection in Fig. 5 is one choice of several, but emphasizes evolutionary trends. On the *Left*, colors resemble Fig. 4 *A–C* to emphasize preexisting ordered code domains. On the *Right*, PR coloring (Fig. 4*C*) emphasizes coexistence of final SGC PR order with distinct columnar wobble domains (Figs. 4*D* and 5).

The Evolution Shown Is One Variant. Fig. 5 describes the transition from an early era of unique codon assignment, before wobble (*Left*), to a late era when Crick wobble was near universal (*Right*). The transition shown is not unique; it could be diagrammed with differing configurations on the *Left* in Fig. 5, yet have SGC-like outcomes on the *Right* (see *The Ordered SGC Was Assembled from Smaller Ordered Parts* below).

Both Difficult and Soft Fitting Can Exist in the SGC. Fig. 3 calculates negative effects of difficult fitting on SGC-like codon assignment. Difficult fitting characterizes regions that cannot overlap, like wobble domains that must preserve their specific amino acid meanings (Fig. 3*A*). However, Fig. 5 illustrates a complementary soft mode of fitting, which increases SGC order. In soft fitting, ordered domains readily coexist, because triplets simultaneously serve more than one role.

To illustrate soft fitting, consider AUU encoding Ile (Fig. 5, *Left*). Ile encoding has multiple associations: the codon AUU is exceptionally frequent within Ile binding sites (25). The anticodon of AUA is also exceptionally frequent. Both are also prominent in maximally probable RNA–Ile binding sites, requiring a minimal number of nucleotides (30). But Ile AUU and AUA are also part of an extended SGC column encoding similar hydrophobic polar requirements, implying both first position (Fig. 4*C*) and third codon position changes (Fig. 4*D*). Moreover, AUU Ile is the terminus of the oxaloacetic acid biosynthetic group (Fig. 4*B*). Ile assignments conserve overlapping stereochemical, chemical, and biosynthetic order simultaneously, because they fit together without conflict.

Less Difficult Fitting May Be Beneficial. Difficult fit, as for wobble domains, unavoidably decreases assignment accuracy, especially when superposed on capture of neighboring unassigned codons (Fig. 3 *B* and *C*). The negative fitting effect on accuracy is ~10-fold for late Crick wobble and high accuracies (Fig. 3*C*). But capture itself is not dispensable. There are many examples of chemical resemblance between metabolically unrelated amino acids assigned to SGC codons one mutation apart, suggesting capture, for example, Phe and Leu (Fig. 3*B*).

Some routes to reduced hard fitting are evident from these data (Fig. 3). Reducing the size of a mutational capture neighborhood reduces its accuracy penalty (Fig. 3 *A–C*). Perhaps the initial definition of capture (1), which supposed that an assigned triplet can capture any unassigned triplet one mutation distant, was too expansive. For example, because transition mutations are usually more frequent than transversions (31), transition preference could define a reduced capture domain (compare in Fig. 3*A*). Because there is an optimally accurate early time for late wobble advent (11), one could favor historical selection of the SGC on the early side of optimal, requiring fewer captures. Certainly, one should not complicate or expand simple Crick wobbles (3).

Unique Amino Acid Assignments Readily Exist before Wobble. Completion of a code (adding 21st and 22nd functions) (1) is uniquely complex for kinetic reasons (Fig. 2 *A–C*). Codon assignment necessarily slows near completion of SGC evolution. Evolution of the last 2 functions, on average, would take much longer and require a more complex set of events than the first 20 encoded functions (Fig. 2).

This is consistent with molecular evidence that definitive initiation and termination were encoded late, after the amino acids, subsequent to separation of life's major domains. For example, translation initiation and termination logic differ in eukarya and bacteria (32, 33) and the protein catalysts involved, necessarily themselves products of a sophisticated translation apparatus, are of independent origin in different domains. Unlike their catalysts, codons for initation/termination are near universal, so shared primordial start/stop mechanisms may have existed also. Nevertheless, primitive but still extant early encoding is probably restricted to that for SGC amino acids.

Difficulty with latter assignments worsens if wobble occurs from the beginning of coding (1, 3), but such barriers can be bypassed by supposing that wobble was instituted late, after the code was substantially formed by unique, nonwobbling assignments. Late wobble is independently plausible, because complex, organized ribosomal conformational changes (5, 34) and a particular anticodon loop conformation (6, 7) and rare base electronic structures (8) are required for accurate wobble pairing. So, before evolution of a complex translation apparatus, standard codon–anticodon base pairing, though perhaps inaccurate, is more plausible than accurate wobble. Moreover, transacylation catalyzed by RNA readily generates a suitable, simpler aminoacyl–tRNA precursor: linear aminoacyl-ribotetramers (35), whose 5′ nucleotides might serve an anticodon-like function (1, 36, 37).

Thus, early coding may still be visible in unique amino acid assignments, without Crick wobble. But the present SGC likely appeared subsequent to late Crick wobble, defined as NNY and NNR translation by individual acceptor RNAs (1). An earlier nonwobbling SGC precursor is implied, with near-complete amino acid assignment. Fig. 5, *Left* shows that this implied, but seemingly improbable, nonwobbling SGC precursor can exist. Further, the nonwobbling precursor is a sufficient foundation for a complete row- and column-ordered SGC (Fig. 5, *Right*).

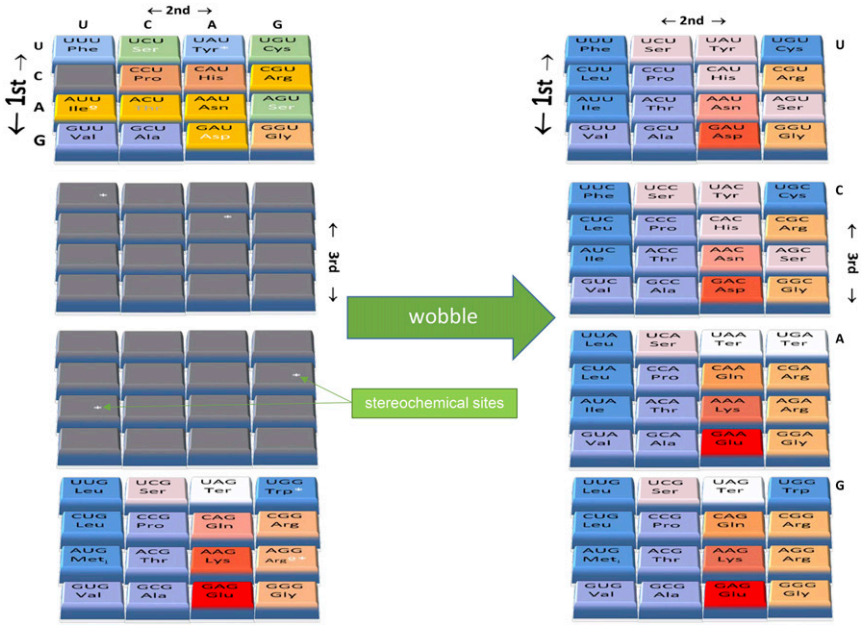

Fig. 5. A unified SGC from three-dimensional late Crick wobble. Three-dimensional coding tables are shown with standard assignments: first nucleotide variation UCAG back to front, second nucleotide variation *Left* to *Right*, and third nucleotide variation *Top* plane to *Bottom* plane in UCAG order. Colors on triplet tiles correspond to those in Fig. 4, explained in the text. Standard three-letter abbreviations for the amino acids identify each codon assignment. Small white symbols are coding triplets exceptionally concentrated in cognate-selected RNA–amino acid binding sites (25). Anticodon concentrations are marked with white *, codon concentrations are marked with white °. The three-dimensional coding table on the *Left* is a possible nonwobble, uniquely assigned, precursor of the complete SGC on the *Right*, where coding after adoption of late Crick wobble is shown.

The Influence of Likely Ancient Amino Acids Is Preserved. A broad consensus finds Val, Ala, Asp, Glu, and Gly to be credible primordial amino acids for reasons of synthetic ease (38), natural abundance (39), and ready thermodynamic access (14). In Fig. 5, this probable primordial status is accepted: all code order ultimately rests on initial encoding of these amino acids by the row of GNN codons in the SGC.

Amino acids probably available before biosynthesis (Fig. 4*A*) occur in both the NNU (*Upper*) plane and the NNG (*Lower*) plane. The *Upper* SGC plane fuses the primordial chemistry row (Fig. 4*A*) with the row-biased biosynthetic domains of Fig. 4*B*. The *Lower* SGC plane combines primordial availability (Fig. 4*A*) with the SGC's chemically organized columns (Fig. 4*C*).

No Recoding Is Required to Add Late Wobble. In Fig. 5, wobble advent requires no changes whatever to initial encoding. This is notable, because wobble as a source of SGC order affects most encoded functions (Fig. 4*D*); its influence on code structure spans the coding table. Wobble has a major evolutionary role in shaping the SGC's close spacing of assignments for related functions (1). Moreover, late wobble is the most probable path to full and complete codes (1, 3). It is noteworthy that a fundamental coding feature can be added after most assignments are made, without perturbing extensive foundations.

This is particularly true in light of negative wobble effects. Larger wobble domains, and earlier wobble also, increase completion complications (Fig. 2). This is more important when larger wobble domains are spread to a triplet's mutational neighborhood by capture of nearby unassigned triplets (Fig. 3). Decreases of more than an order of magnitude in assignment accuracy are routine for use of capture and wobble, with the greatest difficulty occurring for the greatest resemblance to the SGC (Fig. 3 *B* and *C*). Notably, completion complications are entirely relieved via the costless pathway for late-wobble introduction shown in Fig. 5.

The SGC Is Accessible from Experimental Levels of RNA–Amino Acid Stereochemistry. In Fig. 5, *Left* small white symbols on triplet tiles indicate conserved, functional, cognate coding triplets within experimentally selected amino acid binding sites. White symbols indicate eight triplets with nine experimental stereochemical connections, quite varied in amino acid side chain and coding sequences: Ile codon AUU, Tyr anticodon AUA, Phe anticodon GAA, His anticodon GUG, Ile anticodon UAU, Arg anticodon UCG, Trp anticodon CCA, and uniquely: Arg, employing both codon AGG and its anticodon CCU (25).

Nevertheless, How were most SGC sense triplets ordered, beginning from eight or more initial loci? Fig. 5 takes experimental RNA–amino acid interactions as nuclei for small, ordered code substructures. In this way, each *Left*-hand SGC plane has two documented stereochemical nucleations. Thus the SGC, unexpectedly, can appear overdetermined; there are so-far functionless stereochemical sites in unassigned gray areas (Fig. 5, *Left*). Such excess suggests an additional role for amino acid–RNA interactions in onset of Crick wobble, between Fig. 5's *Left* and *Right* halves.

A Stereochemical Role in Wobble Advent. There is a straightforward way, economical of hypotheses, to add unused stereochemical assignments (Fig. 5, *Left*) to late wobble evolution. A two-plane model may be oversimplified, given that there are possible stereochemical triplets and likely ancient Val/Ala/Asp/Glu/Gly assignments on all four SGC planes (Fig. 4*A*). Thus, soft fitting of ordered code precursors might occur through all planes (Fig. 5), utilizing all known amino acid–RNA relations (25). This will bear more thought.

Sources of SGC Order Coexist. Previous analysis (1) suggests that, in the very highly ordered SGC, related assignment to chemically similar (21), related assignment to biosynthetically related amino acids (17) and minimization of amino acid errors (40) exist simultaneously. Further, chemically determined assignments and

error minimization appeared to be statistically independent (41). It has not been clear how such differing molecular objectives could be satisfied together in the SGC. Fig. 5 now shows that different principles can be implemented in separated SGC regions (*Left*) and later combined (*Right*) after accurate wobble evolves.

Selection Acts at the Newly Defined Level. The fit of ordered code parts, simultaneously preserving primordial origins (Fig. 4*A*), biosynthetic relations (Fig. 4*B*), and chemical function (Fig. 4*C*) while allowing wobble coding (Figs. 4*D* and 5) is surely not accidental. Instead, it implies a role for distribution fitness, that is, evolutionary selection of extreme members of a broadly distributed population (1). Via Fig. 5's pathway, selection for efficient translation chose, from a varied population of intermediate codes, unrelated complementing parts that merged into a complete SGC.

This is exemplified by a specific three-dimensional late Crick wobble example. Earlier unique coding might make different assignments to NNA and NNG, or alternatively, to NNU and NNC codons. But using Crick wobble (1), neither NNA nor NNC can be assigned specifically, because later wobble will make them equivalent to NNG and NNU, respectively. However, with freedom to select combined partial codes as in Fig. 5, such conflicts can be avoided; an SGC-like group can be found among an unchanged population of nascent codes.

Late Crick Wobble Precisely Generates Full Codes. Code completion can be variously defined. One might be interested in "full" codes (all triplets assigned) or "complete" codes (all functions encoded). Prior results (1) suggest that late Crick wobble approaches full coding more precisely than continuous wobble, while still allowing coding capacity for later definitive initiation and termination. Three-dimensional late Crick wobble (Fig. 5) now makes superior access to full coding explicit. Late Crick wobble among uniquely assigned precursors (Fig. 5, *Left*) precisely fills the coding table, closely approximating the full three-dimensional SGC (Fig. 5, *Right*) via one uniform transition. This recalls the finding (1) that unassigned triplets readily persist into late code history, and therefore could take late evolutionary roles.

An NNG Intermediate Appears as a Reduced, Capable, Translation System. In Fig. 5, the lower proposed plane encodes 13 amino acids, as well as early initiation and termination. This is particularly striking; a chorismate mutase has been reduced to 14, then to 9 amino acids (42). Reduction of the amino acid complement in nucleoside diphosphate kinases shows that stable structures and enzymatic activities are accessible together if 13 different amino acids are encoded (43). The NNG plane (Fig. 5, *Left*), therefore, is a particularly plausible evolutionary intermediate because it can independently encode accurately terminated, functional enzymes.

An Intricate Development Is Facilitated by Division into Regions. SGC-like codes arising earlier would more probably be selected. Division of SGC encoding into multiple regions allows code evolution in parallel, which potentially shortens time to completion, compared to a single linear path.

This repeats a common theme; evolution in multiple regions is also used to minimize penalties for difficult wobble fit (Fig. 3). Evolution in multiple regions allows simultaneous implementation of varied means to code order (Figs. 4 and 5). Selection of multiple regions acts on smaller, perhaps more readily organized, coding intermediates (Figs. 1 and 5).

Moreover, evolution by division shrinks population size required to select a code. This size is defined by the abundance equation, $S = \ln 2/P_{SGC}$ (11), where S is the number of independent codes that must be examined by selection to find, with probability 0.5, a code occurring with fractional abundance P_{SGC}. However, reassorting preexisting partial codes eventually allows all combinations, diversifying codes available from a fixed number of individuals, thereby shrinking the population required to select an SGC.

The Ordered SGC Was Assembled from Smaller Ordered Parts. The SGC likely originates as less functional partial codes fitted softly to fill a coding table (Fig. 5). Relevant evolution has been studied quantitatively (44). The crucial idea is that newly evolved codon assignments are constrained by history, because new assignments must preserve function of already-encoded peptides. Such conservative selection is sufficient to create coding that approaches SGC levels of order. Such order evolves, even beginning with a code that is uniformly ambiguous (44), so that no relation whatever between codons and amino acids preexists. Ordering works better after a few early assignments, and so is well suited to a few prior stereochemical assignments (Fig. 5, *Left*). It readily produces two-dimensional codes for ~12 amino acids, explicitly supporting planar intermediates like those posited in Fig. 5 (44). Such conservative evolution should therefore transmit an overall chemical logic from initial assignments (25) to subassemblies (compare *AN NNG Intermediate Appears as a Reduced, Capable, Translation System* above), then to the SGC (Fig. 5, *Right*).

Origin of Smaller Ordered SGC Parts. Possibly, ordered SGC subsections evolved in primitive cellular compartments, forming the SGC by compartmental membrane fusions. However, an alternative pathway emulates accepted later events in cellular evolution.

About 2.2 billion years ago, as oxygen initially accumulated in the Earth's atmosphere (45), an α-proteobacterium (46) began an endosymbiotic relationship with a still-uncertain archaeon (47) related to Asgard archaea (48). This bacterium became the ancestor of the eukaryotic mitochondrion, and thus of a chimeric aerobic eukaryote. Chimerism was followed by transfer of numerous genes from endosymbiont to host cell (49). Multicompartment eukaryotic cells notably have distinct genetic codes in different compartments, sometimes relying on transferred tRNA genes (50). A remarkably parallel sequence of events ~1.5 billion years earlier, in which ancient cells with differing, partial codes fused, could have founded the SGC. Vetsigian et al. (51) suggest that horizontal gene transfer created a universal SGC. Horizontal gene transfer, long before universality, could have created the ancestral code (Fig. 5).

Methods

Computed Coding Table Evolutions. Simulations have been described with more detail (1, 11). Calculations begin with an empty coding table. Code evolution is divided into short time slices. In each slice, a nascent coding table is visited by computer. A random triplet is chosen. Each such choice initiates a computed passage, during which only one event: initiation or decay or capture or alternatively, nothing at all, will occur. If the chosen triplet is unassigned, it can be assigned to 1 of 20 amino acids, initiation, or termination. Assignments can be to one triplet (unique), or to a group related by wobble. Subsequent assignment decay also occurs uniquely, or for a wobble group, if such a group exists.

All events happen stochastically, determined by randomized numbers, conferring assigned probabilities for one passage. These procedures are equivalent to assigning chemical rate constants (1), assuming that initiation and decay are first order in unassigned and assigned codons, respectively, and that mutational capture is second order, depending on the product (assigned triplets*unassigned triplets) (1). An important implication is that passages are an appropriate time unit, proportional to real-world time.

It is assumed that representative probabilities for evolutionary events during one passage exist; for example, probability Pinit for initial codon assignment, Pdecay for loss of assignment, Pmut for mutational capture of an unassigned codon one mutation distant [using the coevolution (17)/polar requirement (40) protocol called Coevo_PR (1)] and Pwob for wobble when there is a choice between types of wobble or unique encoding. Prand is the probability that an initial assignment is random, with no specified relation to the same SGC triplet's meaning. Here, unless otherwise specified (for example, when a probability is being varied): Pinit = 0.6, Pdecay = 0.04, Pmut = 0.04, Pwob = 1.0, and Prand = 0.1 for each passage.

Example source code comprising ~900 lines of Pascal, and an Excel spreadsheet example of downstream analysis and graphics, are available at 10.5072/zenodo.733491.

Minimally Evolved Genetic Codes. These have specified numbers of randomly chosen triplets with SGC assignments, complemented by completely random assignments for other codons. Such coding tables may not have mutational capture, and do not evolve in other ways. However, as used here, minimally evolved coding tables do allow prior assignments to decay. Otherwise it can be impossible to complete a coding table that has a requested composition or history. Without decay, such a calculation can hang indefinitely, unable to recover from a poor assignment. In contrast, when decay is possible, difficult evolutions appear instead with a longer completion time, and are readily included in population statistics.

Assignment Accuracy for Fully Random Coding. In Fig. 1*B*, average accuracies are plotted for coding tables filled randomly. These are calculated from the binomial distribution:

$$P(mis) = \frac{asgn!}{(asgn - mis)!mis!}(1 - Prand)^{(asgn - mis)}\, Prand^{mis},$$

where *mis* is the number of misassigned triplets, *P*(*mis*) is the probability of *mis* misassignments when random assignments occur with *Prand* and *asgn* is the total number of assigned triplets. *P*(*mis*) for different *mis* were summed as required to get results in Fig. 1*B*. Fig. 1*A* has a mean *asgn* = 58.56; here this mean term is calculated alone to approximate the distributed occupancies of the complete binomial population. Sums of the natural log of *P*(*mis*) should have an innate dependence on ln (1 − *Prand*), as shown by the above equation and in Fig. 1 *B*–*D*.

Data Availability. Source code and Excel analysis/graphics data have been deposited in Zenodo (DOI: 10.5072/zenodo.733491) (52).

ACKNOWLEDGMENTS. Thanks to Tom Cech for discussion of a draft manuscript.

1. M. Yarus, Evolution of the standard genetic code. *J. Mol. Evol.* **89**, 19–44 (2021).
2. M. Yarus, J. G. Caporaso, R. Knight, Origins of the genetic code: The escaped triplet theory. *Annu. Rev. Biochem.* **74**, 179–198 (2005).
3. M. Yarus, Crick wobble and superwobble in standard genetic code evolution. *J. Mol. Evol.* **89**, 50–61 (2021).
4. D. Moazed, H. F. Noller, Binding of tRNA to the ribosomal A and P sites protects two distinct sets of nucleotides in 16 S rRNA. *J. Mol. Biol.* **211**, 135–145 (1990).
5. J. M. Ogle *et al*., Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science* **292**, 897–902 (2001).
6. M. Yarus, Translational efficiency of transfer RNA's: Uses of an extended anticodon. *Science* **218**, 646–652 (1982).
7. O. C. Uhlenbeck, J. M. Schrader, Evolutionary tuning impacts the design of bacterial tRNAs for the incorporation of unnatural amino acids by ribosomes. *Curr. Opin. Chem. Biol.* **46**, 138–145 (2018).
8. E. Westhof, M. Yusupov, G. Yusupova, The multiple flavors of GoU pairs in RNA. *J. Mol. Recognit.* **32**, e2782 (2019).
9. F. H. Crick, Codon--anticodon pairing: The wobble hypothesis. *J. Mol. Biol.* **19**, 548–555 (1966).
10. S. Alkatib *et al*., The contributions of wobbling and superwobbling to the reading of the genetic code. *PLoS Genet.* **8**, e1003076 (2012).
11. M. Yarus, Optimal evolution of the standard genetic code. *J. Mol. Evol.* **89**, 45–49 10.1007/s00239-020-09984-8. (2021).
12. M. Eigen, W. Gardiner, P. Schuster, R. Winkler-Oswatitsch, The origin of genetic information. *Sci. Am.* **244**, 88–92, 96, et passim (1981).
13. S. L. Miller, A production of amino acids under possible primitive earth conditions. *Science* **117**, 528–529 (1953).
14. P. G. Higgs, R. E. Pudritz, A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology* **9**, 483–490 (2009).
15. F. J. Taylor, D. Coates, The code within the codons. *Biosystems* **22**, 177–187 (1989).
16. P. G. Higgs, A four-column theory for the origin of the genetic code: Tracing the evolutionary pathways that gave rise to an optimized code. *Biol. Direct* **4**, 16 (2009).
17. J. T.-F. Wong, A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci. U.S.A.* **72**, 1909–1912 (1975).
18. T. A. Ronneberg, L. F. Landweber, S. J. Freeland, Testing a biosynthetic theory of the genetic code: Fact or artifact? *Proc. Natl. Acad. Sci. U.S.A.* **97**, 13690–13695 (2000).
19. R. Amirnovin, An analysis of the metabolic theory of the origin of the genetic code. *J. Mol. Evol.* **44**, 473–476 (1997).
20. M. Di Giulio, An extension of the coevolution theory of the origin of the genetic code. *Biol. Direct* **3**, 37 (2008).
21. C. R. Woese, D. H. Dugre, S. A. Dugre, M. Kondo, W. C. Saxinger, On the fundamental nature and evolution of the genetic code. *Cold Spring Harb. Symp. Quant. Biol.* **31**, 723–736 (1966).
22. S. E. Massey, A. Sequential, A sequential "2-1-3" model of genetic code evolution that explains codon constraints. *J. Mol. Evol.* **62**, 809–810 (2006).
23. C. R. Woese, Order in the genetic code. *Proc. Natl. Acad. Sci. U.S.A.* **54**, 71–75 (1965).
24. D. C. Mathew, Z. Luthey-Schulten, On the physical basis of the amino acid polar requirement. *J. Mol. Evol.* **66**, 519–528 (2008).
25. M. Yarus, The genetic code and RNA-amino acid affinities. *Life (Basel)* **7**, 13 (2017).
26. M. Yarus, A specific amino acid binding site composed of RNA. *Science* **240**, 1751–1758 (1988).
27. R. R. Breaker, R. M. Atilho, S. N. Malkowski, J. W. Nelson, M. E. Sherlock, The biology of free guanidine as revealed by riboswitches. *Biochemistry* **56**, 345–347 (2017).
28. L. Bartonek, B. Zagrovic, mRNA/protein sequence complementarity and its determinants: The impact of affinity scales. *PLOS Comput. Biol.* **13**, e1005648 (2017).
29. M. Yarus, J. J. Widmann, R. Knight, RNA-amino acid binding: A stereochemical era for the genetic code. *J. Mol. Evol.* **69**, 406–429 (2009).
30. C. Lozupone, S. Changayil, I. Majerfeld, M. Yarus, Selection of the simplest RNA that binds isoleucine. *RNA* **9**, 1315–1322 (2003).
31. F. Vogel, Non-randomness of base replacement in point mutation. *J. Mol. Evol.* **1**, 334–367 (1972).
32. M. Kozak, Initiation of translation in prokaryotes and eukaryotes. *Gene* **234**, 187–208 (1999).
33. M. V. Rodnina, W. Wintermeyer, Recent mechanistic insights into eukaryotic ribosomes. *Curr. Opin. Cell Biol.* **21**, 435–443 (2009).
34. T. Powers, H. F. Noller, Dominant lethal mutations in a conserved loop in 16S rRNA. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 1042–1046 (1990).
35. N. V. Chumachenko, Y. Novikov, M. Yarus, Rapid and simple ribozymic aminoacylation using three conserved nucleotides. *J. Am. Chem. Soc.* **131**, 5257–5263 (2009).
36. M. Yarus, The meaning of a minuscule ribozyme. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **366**, 2902–2909 (2011).
37. M. Illangasekare, M. Yarus, Small aminoacyl transfer centers at GU within a larger RNA. *RNA Biol.* **9**, 59–66 (2012).
38. S. L. Miller, H. C. Urey, Organic compound synthesis on the primitive earth. *Science* **130**, 245–251 (1959).
39. J. R. Cronin, S. Pizzarello, C. B. Moore, Amino acids in an antarctic carbonaceous chondrite. *Science* **206**, 335–337 (1979).
40. S. J. Freeland, L. D. Hurst, The genetic code is one in a million. *J. Mol. Evol.* **47**, 238–248 (1998).
41. J. G. Caporaso, M. Yarus, R. Knight, Error minimization and coding triplet/binding site associations are independent features of the canonical genetic code. *J. Mol. Evol.* **61**, 597–607 (2005).
42. K. U. Walter, K. Vamvaca, D. Hilvert, An active enzyme constructed from a 9-amino acid alphabet. *J. Biol. Chem.* **280**, 37742–37746 (2005).
43. M. Kimura, S. Akanuma, Reconstruction and characterization of thermally stable and catalytically active proteins comprising an alphabet of ~13 amino acids. *J. Mol. Evol.* **88**, 372–381 (2020).
44. G. Sella, D. H. Ardell, The coevolution of genes and genetic codes: Crick's frozen accident revisited. *J. Mol. Evol.* **63**, 297–313 (2006).
45. J. W. Schopf, Geological evidence of oxygenic photosynthesis and the biotic response to the 2400–2200 ma "Great Oxidation Event". *Biochemistry (Mosc.)* **79**, 165–177 (2014).
46. M. W. Gray, Mitochondrial evolution. *Cold Spring Harb. Perspect. Biol.* **4**, a011403 (2012).
47. J. M. Archibald, Endosymbiosis and eukaryotic cell evolution. *Curr. Biol.* **25**, R911–R921 (2015).
48. A. Spang *et al*., Asgard archaea are the closest prokaryotic relatives of eukaryotes. *PLoS Genet.* **14**, e1007080 (2018).
49. C. Ku *et al*., Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 10139–10146 (2015).
50. T. H. Jukes, S. Osawa, The genetic code in mitochondria and chloroplasts. *Experientia* **46**, 1117–1126 (1990).
51. K. Vetsigian, C. Woese, N. Goldenfeld, Collective evolution and the genetic code. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 10696–10701 (2006).
52. M. Yarus, Pascal source and Excel data for evolution of Standard Genetic Code. *Zenodo*. https://sandbox.zenodo.org/record/733491#.YSaFRudMGMo. Deposited 19 February 2021.