From initial RNA encoding to the Standard Genetic Code

Michael Yarus
Department of Molecular, Cellular and Developmental Biology
University of Colorado Boulder
Boulder, CO 80309-0347

Email: yarus@colorado.edu

Keywords: code fusion, code division, frozen accident, adaptation, error minimization, stereochemistry, Lateral Gene Transfer, ribonucleopeptide, RNP transition, anthology

Abstract

Multiple prior experiments show that RNA binds chemically varied amino acids within specific oligonucleotide sequences. The smallest, simplest, and therefore most likely primitive RNA binding sites frequently contain conserved triplets corresponding to the Standard Genetic Code (SGC). Here, the implications of such cognate coding triplets are calculated, combining them with an optimized kinetic model for SGC evolution. RNA-amino acid interactions at obseved frequencies choose an SGC-like code, and, using the same mechanism, effectively resist alternative triplet assignments. Resistance to other kinds of coding is evident across varied code initiation scenarios. RNA-mediated assignments at experimental frequencies are sufficient to guide the 'ribonucleopeotide (RNP) transition' to a modern RNP code. This can account for extreme selection of the SGC among its astronomical code possibilities; very SGC-like codes are ca. 1/50 to 1/5 of codes within such a population. Nevertheless, full accounting depends on RNA affinities yet unmeasured. Such a code begins as mostly stereochemical, excludes mismatched assignments, and critically relies on properties characteristic of fusible microbes. After its RNP transition in a partially assigned code, evolution accelerates definitively. Other assignment methods (adaptation, co-evolution, revised stereochemistry, LGT) likely complete the modern SGC because stable cellular intermediates with > 1 code exist, allowing compartmental code exchanges. Though initiated using chemical affinities, the 83 order-of-magnitude focus required to find a near-complete SGC among all possible codes was made by sequential evolutionary anthologies, in successive biological settings.

Introduction

An extraordinary bottleneck. Allowing unassigned functions, there are 1.8×10^{83} possible genetic codes with 20 functions and 64 triplets (1). Moreover, available primordial amino acids (2) greatly exceed selected biological forms (3). Carbonaceous meteorites contain at least 75 amino acids, a minority of which are biological (4). Further, it is clear that multiple bacterial codon assignments can be changed without severe biological consequences (5). Therefore, it is most striking that more than 250,000 codes

in current bacterial and archaeal genomes (6) present only variants of the Standard Genetic Code using the same 22 amino acids (7).

Thus, modern biota descend from a single ancestral group. Early life on Earth went through an astronomically extreme bottleneck from which only one lineage, with its Standard Genetic Code, has emerged. Accordingly, much was determined by this event. Consider the choice of encoded amino acids (3, 8); some standard amino acids are inherited from a universal coding ancestor. More completely, the SGC embodies information from a crucial era populated by ancestors of all life on Earth.

Initial coding. Inevitably, numerous amino acids were encoded before the advent of modern aminoacyl-RNA synthetase proteins (aaRS), which acylate RNA adaptors (present tRNAs) with specific amino acids for translation. This follows because aaRS are themselves complex polypeptides, with specific sites for at least three substrates, and specific active site geometries supporting catalyzed regiospecific aminoacyl transfer (9). Thus the aaRS era must have arisen from a simpler type of competent aminoacyl-RNA synthesis and translation, presumably mediated by RNAs (10–12). In particular, RNA-catalyzed amino acid activation (13) and aminoacyl transfer to RNA (11) have been reproduced and studied.

RNA-amino acid interactions. Extensive evidence shows that amino acids are bound specifically to small ribo-oligonucleotide sequences. Though originating with a natural *Tetrahymena* example, abundant other data exist (14–20).

IN particular, RNA binding sites for amino acids have been characterized experimentally. Randomized RNAs were subjected to selection-amplification (21) after affinity chromatography for amino acids (22). Elution with free amino acids ensures that selected RNAs bind them. A summary of 464 such independently-derived sites for 8 chemically-varied amino acids, containing 7137 site nucleotides selected from 21938 initially randomized positions is shown in Table 1 (23). Local sequences within such binding sites contain an excess of cognate coding triplets for their amino acids, when compared to nearby non-binding-site sequences, which came from the same randomized RNA pools and passed through an identical purification. Interestingly, if the smallest binding sequences (the most likely primitive examples) are selected, inclusion of essential coding triplets is complete (24).

Every RNA site is independent because it has different sequences in initially randomized regions outside the binding site itself. In contrast, site sequences are repetitive, because amino acid binding must conserve atomic contacts with ribonucleotides. In fact, such sequence conservation is strong evidence that conserved coding triplets within binding sites serve essential functions. This is confirmed by showing that such triplets cannot be experimentally modified without disrupting binding function (25).

Table 1 links amino acids, their coding triplets and the probability of observed triplet concentrations in RNA binding site nucleotides. Probability cited is that cognate triplets are equally distributed between site and non-site sequences. A low probability threshold corrects for multiple trials (23): in Tab. 1, significant triplet concentrations have probabilities $< 2.3 \times 10^{-4}$.

This analysis of RNA-amino acid binding sites has been questioned (26), but common objections to stereochemical mechanisms are not compelling. Amino acid-ribonucleotide sequence associations do not rely on probability calculations. Site populations consist of highly repeated independent examples of one, or more, families of related site sequences (27). Cognate triplet sequences are directly observed

among binding site nucleotides (identified by conservation and biochemical criteria), so their association cannot be a statistical artifact. Such data do not identify coding triplets as immediate contacts for the amino acids (an early idea tracing to coding in 'holes' on DNA: 28). But contact is not logically required, because conserved proximity of triplet sequences and cognate amino acids is sufficient to explain a later relationship. Moreover, it is sometimes asserted that the SGC mixture of primitive (e.g., Ile) and later amino acids (e.g., Arg) RNA sites is improbable, but it is argued below (Discussion) that this is expected. Similarly, seeking small, prebiotically plausible binding sites commonly mandates moderate dissociation constants (23), but this is not a valid objection to site significance. Michaelis-Menten catalysis imposes rates dependent on forward 1st order velocities, as well as affinities. Whole or fractional millimolar dissociation constants (23) are fully consistent with useful (bio)chemical velocities.

Notable selectivity. Keeping in mind that Tab. 1 surveys only 40% of encoded amino acids, and that modern laboratory conditions may not wholly reproduce the Earth of 4 gigayears ago, observed RNA-amino acid interactions nevertheless suggest potential coding implications.

Generality. Six of eight sufficiently-tested amino acids show potential stereochemical ancestry.

Variety. Varied side chain chemistries exhibit RNA stereochemical links (Tab. 1): aliphatic hydrophobic (Ile), potentially catalytic (His), polar, strong H-bonding (Arg), planar aromatic (Phe, Tyr, Trp).

Specificity. For example, RNA distinguishes similar aliphatic sidechains, isoleucine and leucine (Tab. 1). An L-valine site distinguishes D-valine, L-Leucine, L-isoleucine, and L-norvaline, chemically very similar (29). An RNA binding site selected for phenylalanine can accept other planar aromatics, or reject tyrosine and tryptophan (27).

Triplets. No RNA affinities are associated with amino acid codons alone. All six amino acids sites link amino acids to cognate anticodons, thereby resembling associations in modern translation (30, 31).

RNA-amino acid interactions provide a unique code origin. Above, the astronomically vast universe of possible codes is emphasized. Concentration on one possibility of $\approx 10^{83}$ must be explained.

Thus an effective, specific mechanism for bringing RNA sequences together with amino acids is required. Such a mechanism almost identifies itself. It is difficult to imagine encoding without a physico-chemical interaction inducing proximity between specific amino acids and cognate coding sequences.

Thus, stereochemical origins are unique for two reasons. First, because of the need for choice among an astronomical code variety. Second, to juxtapose reactants and so (at least) entropically (32) catalyzing encoded joining reactions.

Frances Crick supposed that the SGC might be a "frozen accident", and "in its extreme form", coding assignments could all be due to chance (31). But the SGC is not random (23, 33). But once evolved, the code would not be resistant to change, because of the necessity to preserve pre-existing essential gene products (34). However, Crick anticipated that some amino acids might be initially encoded

stereochemically (31). The finding that six of eight amino acids, of varied types, are related to cognate triplets (Tab. 1) suggests that chemical determinism rather than undirected chance has the larger role.

It is also possible that the code is 'adaptive', in the sense that it was selected to minimize phenotypic error when a coding mistake is made (35). Adaptive code evolution implies pre-existing coding whose assignments will be guided toward minimal error. So, adaptation does not initiate coding, nor choose among $\approx 10^{83}$ possibilities.

Evolution of the code might also be shaped by 'coevolution' with amino acid synthesis, with late-evolving amino acids also encoded later (36, 37). Coevolution is a persuasive idea, but offers no way to initiate coding, implement molecular association, or to make astronomically specific choices.

Given that the only apparent way to realize effective coding is with specific stereochemical interaction (38, 39), a mechanistic relation (below) between chemical and other code-ordering mechanisms will clarify code evolution.

Chemistry for a code. Presumably cognate codons, anticodons, or both proximal to an amino acid binding site can be counted as stereochemical links, useable in forming a later code based on binding (23, 40). By this criterion, 6 of 8 examined amino acids have such a link (Tab. 1). Arg, His, Ile, Phe, Trp and Phe show evidence of stereochemically-originated coding, while no such evidence appears for Gln or Leu (23). In Tab. 1, 6 of 8 amino acids still show a stereochemical link if only anticodons are taken as significant (41).

Thus nine cognate SGC triplets are significantly present in RNA binding sequences for the six amino acids. However, CCU 3'/AAG 3' are complementary anticodon/codons for Arg, so imply only one stereochemical assignment. UAU 3'/AUU 3' are also both cognate for Ile, but are not complementary and so are separate chemically-determined assignments. So: 6 of 8 amino acids are putatively related to 8 SGC triplets. We will call this set of assignments "Stereo" to concisely refer to Tab. 1 chemical data. While extrapolation is not unique, extrapolation to 20 amino acids would implicate (20/8) x 8 = 20 stereochemical triplet assignments for 20 SGC amino acids. Thus, on simple assumptions, $\approx$ (20/61) $\approx$ 1/3 of amino acid-encoding triplets in the near-universal genetic code arguably arose from interactions between amino acids and ribonucleotide sequences. Extensions of observed 8 stereochemical triplets to complete coding tables are called _resi1 through _resi5 below, for ***r***andom ***e***xtension of ***s***tereochemical ***i***nitiation (supplementary data).

A likely mechanism for coding table evolution. The path of least selection provides the most probable evolutionary route for a favored characteristic (42). Least selection for genetic encoding utilizes fusion of separate partial codes (43). Such code fragments might have been selected to serve varied functions, so fusing codes have continually increasing biological abilities. Increasing functionality makes fusions easily selected. In addition, code fusions implement evolution in parallel, so yielding complete codes more quickly (43). Given a prominent role for code fusions (44), either homogeneous or heterogeneous fusions leading to the SGC appear possible (45). That is, fusions may join further-evolved descendants of one original code, or may join codes from independent origins.

The purpose. Below, amino acid-RNA associations are set within the shortest, most accurate code evolution known, to study their joint outcomes.

Results

Early amino acid-RNA interactions favor similar later code expansion. In Figure 1, code evolution environments initiate very differently: with either a single SGC-like assignment, or with a complete stereochemical SGC complement, as in **Chemistry for a code** above. Initial codes are completed (evolve to >= 20 assigned functions) by adding single random SGC-like assignments with differing accuracies (having a probability Prand of randomness, e.g., without relation to the SGC).

Fig. 1 plots frequencies of the most SGC-like final codes, having >= 20 encoded functions with no coding differences from the SGC (mis0) or with one non-SGC assignment (mis1). Upper Fig. 1 curves are for 20 initial stereochemically determined triplets, compared to lower curves for single-codon initial assignment.

Large differences result from distinct beginnings; initiation of code evolution with observed levels of stereochemical assignment makes completion of the code much more tolerant to error (randomness in later assignments, Prand). Thus, later assignments can contain errors, while still evolving to approach the SGC. Stereochemical initiation gathers other compatible assignments to itself, allowing later divergence among assignments without making SGC-like coding inaccessibly rare. These effects are large; more-than-order-of-magnitude differences in final code abundances occur even for moderately greater random coding (Fig. 1).

Amino acid-RNA interactions reduce acceptability of divergent assignments. Just above, initial SGC-like stereochemical assignment favors later similar SGC-like assignment. There is a complementary effect on non-SGC assignments, disfavoring infusion of later, conflicting assignments.

Fig. 2 compares incorporation from an entirely different coding scheme (ALT, supplementary data) into near-complete SGC codes, using three kinds of initial coding. Fig. 2A shows data for single random initiations, then continuation with single random SGC assignments. Fig. 2B shows initiation with the 8 experimentally implicated triplets of Table 1, then single random. Fig. 2C shows codes initiated with 20 stereochemically implied initiations appropriate to a complete coding table (**Chemistry for a code**, above), then single random. For each initiation, incorporation of an alternative code that has no assignments in common with the SGC (ALT) into the total population of codes (labeled **pop**) is compared to incorporation into the most complete code (labeled **code**).

All panels share a Y-axis. Thus increasing initial stereochemistry, from single SGC assignments, to Stereo, to _resi2, decreases alternative coding (Fig. 2A, 2B, 2C). Though this has a simple basis (increasing numbers of codons are committed early), it is still a notable effect. For example, the average rightward code in Fig. 2A has seen 76 alternative assignments 5 at a time, but only 3.8 will be accepted in a superior SGC-like (>= 20 function mis 0) code.

Fig. 2 also distinguishes ALT incorporation into the entire code population (**pop**) from that into the most complete codes (**code**) for 1 ALT codon/assignment attempt to 5 codons per patch. Codes formed from single SGC triplets accept more alternative coding than exists in the general code population, thus using alternative assignments to complete their encoding (the arrow relates population mean to most complete codes, Fig. 2A). Coding restricted to experimental amino acid-triplet association (Fig. 2B) is intermediate, actually crossing over at 3 triplet patches. These intermediate initiations finally reject alien assignments, yielding most complete codes with fewer foreign codons (downward arrow) than the code population. Finally, initial assignment with stereochemistry calculated for total amino-acid-RNA sites ((23), Table 1, Fig. 2C) discriminates strongly against alternative assignments, with fewer alternative encodings in most-complete codes at all ALT patch sizes (downward arrow, Fig. 2C).

So: alternative coding in a nascent SGC-like code is limited both with and without a body of early stereochemical assignments, but alternative encoding is particularly rejected with full stereochemical initiation, which accepts a mean of < 1 foreign assignment.

Full stereochemical initiations differ greatly. For clarity, stereochemical initiations have been introduced as if they were uniform. But they emphatically are not. Code extension _resi2 (supplementary data) was used for illustrations (Fig. 1, 2), but other random extensions of known amino acid-RNA interactions have distinct outcomes. This is exemplified in Fig. 3A, which shows the mean time in passages required to evolve near-complete coding in 1500 evolutionary environments. When evolution of a stereochemically-initiated complete code requires more time, that code also admits proportionately more assignments from an independent coding system. Measurements show 8 triplets from evolved RNA binding sites (Stereo, Tab. 1, **Chemistry for a complete code)**, as well as 5 different extensions to completed codes (supplementary data). Time to complete a code varies 6.2-fold among stereochemical initiations, or 3.4–fold among the five completed initial triplet sets.

Stereochemical initiation differences have a simple rationale. Stereochemical initiations differ because of differing numbers of functions encoded. Time to evolve complete codes (Fig. 3A) and other evolutionary indices will vary with evolutionary distance to the SGC, in accord with a principle of least selection (42). Because full initial encodings are derived by adding randomly-chosen SGC triplets for non-studied amino acids to the 8 experimentally determined stereochemical triplets (Tab. 1), prospective full initiations based on the same observed 8 initial triplets differ in number of final functions encoded.

This is reflected in Fig. 3B, where the number of admitted ALT assignments is linearly related to the number of initiating encoded functions. The linearity extrapolates plausibly; with >= 20 functions initially (rightward in Fig. 3B), there is little or no room for alternative coding.

Such initial variation among stereochemical sets of triplets is yet more dramatic if a more informative distance to the SGC (45) is plotted. Such a distance measure reflects both amount of change and evolutionary accuracy. In Fig. 3B, the fraction of >= 20 function codes that are identical to the SGC (mis0 codes), starting with different initiations, is used. About 19% of all >= 20 function codes are identical to the SGC after _resi1 initiation (rightward, Fig. 3B). Observation of almost a fifth of near-complete codes identical to the SGC (Fig. 3B) concisely evokes the power of stereochemical initiation. It should be simple to select an SGC-like code among such an evolving code population. In particular, it appears that the challenge of a single SGC on Earth from 1.8×10^{83} potential coding tables can be readily met via selection

from code fusion-division-late-wobble evolution, initiated with favorable stereochemical assignments. Alternatively: 19% of such nascent codes span the 83 order-of-magnitude SGC focus. Even for the least effective extensions, _resi3, _resi4 and _resi5, least SGC selection seeks ≈ 1 code in 50, still a seemingly short gap and simple task.

Resisting inaccurate assignment. In Fig. 2 and Fig. 3, later code assignments are 15% random (Figure legends). These evolving code populations therefore benefit from Fig. 1's effect. They find the SGC even when challenged by later departure from specific assignment, which would otherwise make an SGC inconveniently rare.

Code populations reside in fusible microbes. Above results, in which the SGC is found despite somewhat unfavorable conditions, are due in part to exploitation of distribution fitness (1); selecting a favorable, less probable minority among a large, less-fit population. Fig. 2 populations range up to 10^5 individual codes (Methods). Besides being numerous, they are able to fuse codes readily by fusing their coding compartments. Both conditions are readily rationalized if they are fusible unicompartmental microbes. Moreover, numerous microbes are also required to exploit the upper tail of distributed codes; that is, evolving by truncation, accepting only atypical best examples (42).

Discussion

Optimized evolution of coding tables. We draw the implications of observed stereochemistry using a Monte Carlo kinetic model of code evolution (1). The model has been optimized by determining the most rapid and accurate of 32 possible pathways for SGC evolution (45). Calculation constructs codes assuming that code evolution obeys standard kinetic rules (1). Different evolutionary pathways of can be compared using a principle of least selection (42), which asserts that the code most rapidly evolved and adhering most closely to known SGC assignments will require least selection; that is, will supply the most probable ancestors of the historical SGC.

Selected major conclusions of prior analysis are that only broadly validated evolutionary means are required to reach the SGC (45), that wobble was likely a later coding development, after most elementary assignments (1, 45, 46), that such wobble was more likely to have been simplified Crick wobble (45) than four-fold superwobble (47, 48), that the SGC was constructed by fusion of separately-evolved coding fragments (43), and that code division and re-fusion of evolved fragments was a major factor, notably enabling SGC evolution from one primary source of code assignments (45). Because fusing codes become unfit if they attribute ambiguous meanings to assigned triplets, fusion pathways purify coding, tending to converge on an underlying coding consensus, like the SGC (44). Moreover, fusion produces a prolonged elevation, or "crescendo" of changing codes, all very SGC-like. Within such a crescendo, SGC can be found by leisurely selection among as many competent codes as required (44, 45).

Code fusion implies code convergence. Fig. 4A displays two merging code sources; if they use differing coding assignments, fused codes assign triplets ambiguously. This makes such codes less fit. Such ambiguous fusions become extinct. A minority of codes surviving fusion therefore has consistent

assignments. That is, surviving codes converge to a consensus. This supplies the likely route to a coherent SGC (44). The same mechanism rejects completely independent codes (Fig. 2).

Such discrimination is more strict for complete stereochemistry (Fig. 2C) than if restricted to experimentally confirmed effects (Fig. 2B), and is least exclusionary in the extreme case that random SGC codons are added at random times, one-by-one (Fig. 2A) instead of being added in an initial set (Fig. 2B, 2C). If SGC codons are added one-by-one, the most complete coding tables actually import a small number of alternative assignments to complete themselves (Fig. 2A). In contrast, initial full-code stereochemical assignments almost exclude alternative assignments, accepting less than one codon per near-complete code (Fig. 2C).

But full stereochemical initiations differ. Completed sets of potential stereochemical assignments differ greatly among themselves (Fig. 3, supplementary data). These differences affect most properties of interest; but 3.4-fold in time to evolve near-complete codes, and 6.5-fold in the number of alternative assignments admitted (Fig. 3A). This is due to varied initial coverage of the SGC: differing sets of triplets with the same stereochemical assignment frequency nevertheless initially assign different fractions of total SGC functions. The evolutionary path from such initiations to the SGC therefore has different lengths, requires varied times (Fig. 3A), and allows different proportions of unlike assignment (Fig. 3B). This conclusion does not require extrapolation: even prior known stereochemistry limits alien assignment to less than 2 codons (Fig. 2B) of the 64 possible in full codes. A persuasive estimate of alternative coding within the SGC will require a completed stereochemical census (23) for SGC triplets.

Magnitude of the RNP transition. The RNP transition solves an acute problem. Early coding must be substantially complete in order to encode initial peptide aaRS, peptide factors, and other proteins required for near-modern protein synthesis. Fortunately, minimal amino acid complexity required for functional proteins has been extensively studied. Tertiary folding is the simpler requirement, and a 3-fold symmetric 140-aa fibroblast growth factor folds well when composed of 12 or 13 amino acids (50), rather than the standard complement.

Enzymatic folded forms are a more complex and relevant requirement for the current argument. Chorismate mutase, a 93-amino acid enzyme with a straightforward alpha-helical structure, has in vivo activity if reduced to only 9 amino acids (51). A 213-aa orotidine phosphotransferase (from PRPP) is active if reduced to 13 kinds of amino acids (52). A reconstructed ancestral bacterial NDP kinase functions with excellent thermal stability and up to 1% control activity with 13 or 14 amino acid alphabets (53). A protein of particular relevance, small ribosomal subunit protein uS8, with 130-aa, binds its RNA if composed of 15, 14 or 13 amino acids (54). Notably, though less stable than true thermophilic homologues, reduced uS8 proteins still have high stabilities, half-melted at ~85°C.

In short, 9 to 15 different encoded amino acids can construct structurally and chemically functional peptides hundreds of amino acids long, with a frequent threshold at 13 different amino acids. The extrapolated stereochemical encoded complement of 15 amino acids (**Chemistry for a code**, above) therefore suggests that RNA can encode the RNP transition.

Thus an initial code relying on RNA chemistry alone probably encoded polypeptides with major structural and enzymatic capabilities (Fig. 4B). After the resulting ribonucleopeptide (RNP) transition,

evolution to modern translation and the full SGC is enabled, requiring exquisite adaptation between proteins and RNAs (55) and wide-ranging ribonucleotide refinement for coding (56, 57), but not radical evolutionary innovation.

Life's tempo and mode after the RNP transition. A characteristic of RNA catalysis is that, though certainly more versatile and faster than once imagined, it is still limited and slower, compared to protein catalysis (58). Biocatalysis will accelerate by orders of magnitude after the RNP transition brings peptide groups to bear on (bio)chemical transition states. The addition of histidine to coding (59, 60) and peptides (61) will likely, even alone, have a notable catalytic impact. As a result, cell activities, particularly including cell division, accelerate through the RNP transition. Evolutionary success is almost synonymous with faster growth (62); thus the definitive early code plausibly emerges from post-RNP cell divisions. RNP transitioning cells race away from those with slower, less versatile catalysis, and we are their inheritors. Such a path, with an effective stereochemical code formed early, accommodates fusion among descendants of a single code ancestor, allowing both a rapid origin and least selection (45).

DNA appears after the RNP transition. Deoxyribose is not among easily accessible 5-carbon sugars (63) and the unusual radical catalysis universally required to generate it from ribose (64) would require biosynthesis of a sophisticated catalyst. This indicates that the DNA genome appears substantially after the RNP transition, as shown in Fig. 4B.

In fact, consolidation of the DNA genome and thus Darwinian descent itself might have depended on pre-existing peptide encoding/translation (42) and thus on the RNP transition. Effective post-RNP translation of numerous encoded peptides sharply elevated the benefits of accurate inheritance.

Fate of early assignments after the RNP transition. Crick's intuition about the 'freezing' of the code (31) applies to the RNP transition (Fig. 4). Even with dramatic changes attending the appearance of peptidyl aaRS and ribosomes, RNA assignments would likely be mostly preserved in the early RNP era to continue synthesis of critical peptides (65). So early triplet assignments likely survive in the SGC, though the machinery selecting them has changed. Separating RNA-era assignments and more complex RNP assignments is clearly worth specific thought.

For example, amino acids assigned multiple sets of codons may have been assigned at two times, in two different ways. Accordingly, positing that Table 1 presents RNA-era encodings is tempting. But this is <u>not</u> plausible; the truth must be more intricate. For example, arginine has two sets of codons - CGN and AGR, has a complex biosynthesis (66), probably was encoded later (67). Arg exhibits multiple affinities to cognate RNA sequences (14, 15, 23, 68). Some Arg assignments can reflect RNA stereochemistry used for assignment after Arg biosynthesis became efficient in the RNP-era, while other assignments reflect multi-molecular affinities, possible only in the RNP-era. Accordingly, even complex amino acids, with late biosynthesis, can be encoded via cognate RNA affinity. However, genuinely persistent RNA-era assignments can exist (consider Ile, Tab. 1), as well as assignments requiring post-RNP molecular collaborations and biosynthesis. More study may disentangle possible encoding histories.

The SGC probably does not have a single origin. Present calculations therefore imply that there was an initial era almost entirely stereochemical, then a second era permitting other kinds of amino acid/triplet assignment. It thereby explains why a metabolic code origin (36) cannot be complete (69, 70) why a

completely adaptive origin (35) becomes difficult to rationalize (71), why a random origin is improbable (72), and indeed, why even a stereochemical origin (23) will not explain every coding assignment (73).

Coding is partial at the RNP transition. Above, coding is incomplete at the RNP transition, likely biased toward early and RNA-interacting amino acids. How is the SGC completed?

Code logic can be more accepting after the RNP transition. Abstractly, post-RNP coding might continue to reject foreign assignments via fusion incompatibility. But an appealing hypothesis is: as the biome multiplied, endocytotic cellular fusions became more probable. Differing codes would coexist in one cell, in different membrane-bound compartments. Such persistent proximity between divergent codes, post-RNP transition, can elevate the frequency of successful code fusion via LGT.

Somewhat similar later events are well-studied. During the DNA era (Fig. 4B), massive multi-directional genetic transfer between compartments was the rule. For example, extensive LGT occurred from the alpha-proteobacterial precursor of the mitochondrion to the protoeukarotic nucleus (74). Only 10-16% of the modern mitochondrial proteome appears to persist from its bacterial ancestor, and at least 842 different initial protobacterial gene families have been identified in other modern cellular locations, including the cell nucleus (75). Gene transfer between DNA compartments was evidently strongly favored. This could augment the code, as well as promoting other evolutionary joinings.

The RNP transition in coding history. Fig. 4B puts the RNA era in historical context, specifically respecting the intricate history of genetic coding. First letters of thhe figure's text entries mark time-of-origin for various events.

From left to right: an RNA world requires the ability to replicate RNAs accurately, though eased by the finding that active translational RNAs can be as small as tetramers, pentamers and hexamers (76–78). Initial coding assignments are made using RNA affinities alone (23, 40). Short, simple encoded peptides can be made on an RNA proto-ribosome, before a fully functional RNP ribosome exists (79). As peptides proliferate, they can join with ribonucleotides to add to the specificity of amino acid binding and triplet assignment (80, 81). The possibility that other ancient molecules also joined RNA to form later amino acid binding sites should not be neglected: polar lipids (82–85) and condensates (86) mediate other well-characterized RNA associations.

Early assembly of the code relies on the code division-fusion crescendo which helps the code converge to the SGC, also speeding its evolution by allowing parallel progress and presenting a prolonged series of close SGC relatives, facilitating the code's optimal least selection (43–45). A set of assignments results, more-than-sufficiently complete to make the RNP transition (shaded region, Fig. 4B) to complete coding. RNA-peptide collaboration in coding functions appears during the RNAP transition, yielding wobble (46) and initiation and termination. Translation initiation (87) and termination (88) change during both early and late SGC evolution, evolving over a prolonged period that continues after the separation of major biological domains. Wobble is a late event during triplet assignment, because accurate wobble requires a large allosteric ribosome (89) and a refined tRNA structure (90, 91). Additionally, early wobble specifically and greatly hinders code completion (46).

After the RNP transition, accurate coding of substantial peptides enables the appearance of an early but recognizably modern ribosome, usually modeled as the predecessor of the modern peptidyl transferase

region (92, 93). Near-modern coding capabilities produce the protein aaRS (94), and deoxyribonucleotides (64). Thus, DNA, the genome and Darwinian evolution arise. Accordingly, the Initial Darwinian Ancestor emerges (95). These events are accompanied by nucleotide modifying enzymes that refine the codon-anticodon complex (56, 57). One complete code becomes the universal SGC (96) and diverges from LUCA to produce a modern multi-domain biota employing the Standard Genetic Code (97).

Biology as Anthology. This adage defines life via its ability to combine advantages evolved separately (45). Here a prior ambiguity is resolved (45): efficient code division-and-fusion SGC evolution defines a unified route to the final code, but also seems to offer code fusion from varied origins. Fusion into a nascent SGC is possible but strictly limited, decreasing as chemically-based initial assignment increases (Fig. 1, 2, 3). Thus, code fusion suggests, but also tenaciously limits, early code joining (44).

The historical SGC is thus a constrained code anthology. It still bears evidence of early stereochemical sources, conserved because assignments must be substantially conserved immediately after the RNP transition to preserve prior peptide capabilities. But early non-stereochemical RNA-era patches are small or nonexistent, dependent on the precise stereochemical history of the code (Fig. 2, 3, 4). The SGC also probably accepted post-RNP transition assignments, because coding was incomplete at the RNP transition. Later, non-stereochemical contributions remain to be resolved, but this can be clarified by practical experiments.

Biology as Anthology; sources of a 10^{83}-fold convergence. In the RNA world, code fusion combining RNAs making specific chemical contacts with amino acids defined, restricted and homogenized a partial, but competent, code, possibly descended from a single ancestor (45). The most complete fused RNA codes supported a decisive RNAP transition, which speedily established its superiority. Then, in the post-RNP era, a different code fusion mechanism added to coding, adapting it to more competent, complex cells. In the DNA era, accurate, near-complete codes competed: a broadly-adapted victor became the SGC (96), efficiently encoding LUCA's diverse proteins (98). The SGC emerges from three sequential anthologies: first joining the reactions of RNAs (23), then joining the competence in coding compartments (43), then joining proficiencies within populations (96).

References

1. M. Yarus, Evolution of the Standard Genetic Code. *J. Mol. Evol.* **89**, 19–44 (2021).

2. S. L. Miller, Production of amino acids under possible primitive earth conditions. *Science* **117**, 528–529 (1953).

3. H. J. Cleaves, The origin of the biologically coded amino acids. *J. Theor. Biol.* **263**, 490–498 (2010).

4. S. Pizzarello, The Chemistry of Life's Origin: A Carbonaceous Meteorite Perspective. *Acc. Chem. Res.* **39**, 231–237 (2006).

5. N. Ostrov, *et al.*, Design, synthesis, and testing toward a 57-codon genome. *Science* **353**, 819–822 (2016).

6. Y. Shulgina, S. R. Eddy, A computational screen for alternative genetic codes in over 250,000 genomes. *eLife* **10**, e71402 (2021).

7. J. Yuan, *et al.*, Distinct genetic code expansion strategies for selenocysteine and pyrrolysine are reflected in different aminoacyl-tRNA formation systems. *FEBS Lett.* **584**, 342–349 (2010).

8. A. L. Weber, S. L. Miller, Reasons for the Occurrence of the Twenty Coded Protein Amino Acids. *J Mol Evol* **17**, 273–284 (1981).

9. M. A. Rubio Gomez, M. Ibba, Aminoacyl-tRNA synthetases. *RNA N. Y. N* **26**, 910–936 (2020).

10. W. Gilbert, The RNA world. *Nature* **319**, 618 (1986).

11. M. Illangasekare, M. Yarus, Small aminoacyl transfer centers at GU within a larger RNA. *Rna Biol.* **9**, 59–66 (2012).

12. T. J. Wilson, D. M. J. Lilley, The potential versatility of RNA catalysis. *Wiley Interdiscip. Rev. RNA* **12**, e1651 (2021).

13. R. K. Kumar, M. Yarus, RNA-catalyzed amino acid activation. *Biochemistry* **40**, 6998-7004. (2001).

14. M. Yarus, A specific amino acid binding site composed of RNA. *Science* **240**, 1751–1758 (1988).

15. M. Yarus, E. L. Christian, Genetic Code Origins. *Nature* **342**, 349–350 (1989).

16. D. B. Johnson, L. Wang, Imprints of the genetic code in the ribosome. *Proc Natl Acad Sci USA* **107**, 8298–8303 (2010).

17. A. S. Rodin, E. Szathmáry, S. N. Rodin, On origin of genetic code and tRNA before translation. *Biol. Direct* **6**, 14 (2011).

18. A. Sengupta, H.-L. Sung, D. J. Nesbitt, Amino Acid Specific Effects on RNA Tertiary Interactions: Single-Molecule Kinetic and Thermodynamic Studies. *J. Phys. Chem. B* (2016).

19. T. H. Kapral, F. Farnhammer, W. Zhao, Z. J. Lu, B. Zagrovic, Widespread autogenous mRNA-protein interactions detected by CLIP-seq. *Nucleic Acids Res.* **50**, 9984–9999 (2022).

20. A. Marton Menendez, D. J. Nesbitt, Lysine-Dependent Entropy Effects in the B. subtilis Lysine Riboswitch: Insights from Single-Molecule Thermodynamic Studies. *J. Phys. Chem. B* **126**, 69–79 (2022).

21. D. Irvine, C. Tuerk, L. Gold, Selexion. *J. Mol. Biol.* **222**, 739–761 (1991).

22. J. Ciesiolka, *et al.*, Affinity selection-amplification from randomized ribooligonucleotide pools. *Methods Enzym.* **267**, 315–35 (1996).

23. M. Yarus, The Genetic Code and RNA-Amino Acid Affinities. *Life* **7**, 13 (2017).

24. C. Lozupone, S. Changayil, I. Majerfeld, M. Yarus, Selection of the simplest RNA that binds isoleucine. *Rna* **9**, 1315–1322 (2003).

25. I. Majerfeld, M. Yarus, A diminutive and specific RNA binding site for L-tryptophan. *Nucleic Acids Res* **33**, 5482–93 (2005).

26. E. V. Koonin, A. S. Novozhilov, Origin and evolution of the genetic code: The universal enigma. *IUBMB Life* **61**, 99–111 (2008).

27. M. Illangasekare, M. Yarus, Phenylalanine-binding RNAs and genetic code evolution. *J Mol Evol* **54**, 298–311 (2002).

28. G. Gamow, Possible Relation between Deoxyribonucleic Acid and Protein Structures. *Nature* **173**, 318–318 (1954).

29. I. Majerfeld, M. Yarus, An RNA pocket for an aliphatic hydrophobe. *Nat Struct Biol* **1**, 287–292 (1994).

30. R. W. Holley, *et al.*, Structure of a ribonucleic acid sequence. *Science* **147**, 1462–1465 (1966).

31. F. H. C. Crick, The Origin of the Genetic Code. *J Mol Biol* **38**, 367–379 (1968).

32. A. Sievers, M. Beringer, M. V. Rodnina, R. Wolfenden, The ribosome as an entropy trap. *Proc Natl Acad Sci U A* **101**, 7897–901 (2004).

33. C. R. Woese, Order in the genetic code. *Proc Natl Acad Sci USA* **54**, 71–75 (1965).

34. G. Sella, D. H. Ardell, The coevolution of genes and genetic codes: Crick's frozen accident revisited. *J. Mol. Evol.* **63**, 297–313 (2006).

35. S. J. Freeland, L. D. Hurst, The genetic code is one in a million. *J Mol Evol* **47**, 238–248 (1998).

36. J. T.-F. Wong, A Co-Evolution Theory of the Genetic Code. *Proc Natl Acad Sci USA* **72**, 1909–1912 (1975).

37. M. Di Giulio, An extension of the coevolution theory of the origin of the genetic code. *Biol. Direct* **3**, 37 (2008).

38. M. Yarus, Amino Acids as RNA Ligands: a Direct-RNA-Template Theory for the Code's Origin. *J Mol Evol* **47**, 109–117 (1998).

39. W. Ma, The scenario on the origin of translation in the RNA world: in principle of replication parsimony. *Biol Direct* **5**, 65 (2010).

40. M. Yarus, J. G. Caporaso, R. Knight, Origins of the genetic code: the escaped triplet theory. *Annu Rev Biochem* **74**, 179–98 (2005).

41. A. L. Weber, J. C. Lacey Jr, Genetic Code Correlations: Amino Acids and Their Anticodon Nucleotides. *J Mol Evol* **11**, 199–210 (1978).

42. M. Yarus, Evolution and change: a principle of least selection. 2021.06.27.450095 (2022).

43. M. Yarus, Fitting the standard genetic code into its triplet table. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2021103118 (2021).

44. M. Yarus, A crescendo of competent coding (c3) contains the Standard Genetic Code. *RNA N. Y. N*, rna.079275.122 (2022).

45. M. Yarus, The Genetic Code Assembles via Division and Fusion, Basic Cellular Events. *Life Basel Switz.* **13**, 2069 (2023).

46. M. Yarus, Ordering events in a developing genetic code. *RNA Biol.* **21**, 1–8 (2024).

47. S. Alkatib, *et al.*, The Contributions of Wobbling and Superwobbling to the Reading of the Genetic Code. *PLOS Genet.* **8**, e1003076 (2012).

48. M. Yarus, Optimal Evolution of the Standard Genetic Code. *J. Mol. Evol.* **89**, 45–49 (2021).

49. M. Yarus, Order of events in a developing genetic code. 2022.12.31.522385 (2023).

50. L. M. Longo, J. Lee, M. Blaber, Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 2135–2139 (2013).

51. K. U. Walter, K. Vamvaca, D. Hilvert, An active enzyme constructed from a 9-amino acid alphabet. *J. Biol. Chem.* **280**, 37742–37746 (2005).

52. S. Akanuma, T. Kigawa, S. Yokoyama, Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. *Proc. Natl. Acad. Sci.* **99**, 13549–13553 (2002).

53. R. Shibue, *et al.*, Comprehensive reduction of amino acid set in a protein suggests the importance of prebiotic amino acids for stable proteins. *Sci. Rep.* **8**, 1227 (2018).

54. F. Zhao, S. Akanuma, Ancestral Sequence Reconstruction of the Ribosomal Protein uS8 and Reduction of Amino Acid Usage to a Smaller Alphabet. *J. Mol. Evol.* **91**, 10–23 (2023).

55. 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).

56. H. Grosjean, E. Westhof, An integrated, structure- and energy-based view of the genetic code. *Nucleic Acids Res.* **44**, 8020–8040 (2016).

57. L. Lei, Z. F. Burton, “Superwobbling” and tRNA-34 Wobble and tRNA-37 Anticodon Loop Modifications in Evolution and Devolution of the Genetic Code. *Life Basel Switz.* **12**, 252 (2022).

58. M. Yarus, Primordial Genetics: Phenotype of the ribocyte. *Annu. Rev. Genet.* **36**, 125–151 (2002).

59. I. Majerfeld, D. Puthenvedu, M. Yarus, RNA affinity for molecular L-histidine; genetic code origins. *J Mol Evol* **61**, 226–35 (2005).

60. J. Ruta, *et al.*, Chiral resolution of histidine using an anti-D-histidine L-RNA aptamer microbore column. *J Chromatogr B Anal. Technol Biomed Life Sci* **845**, 186–90 (2007).

61. R. M. Turk-Macleod, D. Puthenvedu, I. Majerfeld, M. Yarus, The plausibility of RNA-templated peptides: simultaneous RNA affinity for adjacent peptide side chains. *J Mol Evol* **74**, 217–25 (2012).

62. R. E. Lenski, *et al.*, Evolution of competitive fitness in experimental populations of E. coli: what makes one genotype a better competitor than another? *Antonie Van Leeuwenhoek* **73**, 35–47 (1998).

63. Y. Furukawa, *et al.*, Extraterrestrial ribose and other sugars in primitive meteorites. *Proc. Natl. Acad. Sci.* **116**, 24440–24445 (2019).

64. A. M. Poole, *et al.*, The Case for an Early Biological Origin of DNA. *J. Mol. Evol.* **79**, 204–212 (2014).

65. D. H. Ardell, G. Sella, No accident: genetic codes freeze in error-correcting patterns of the standard genetic code. *Philos. Trans. R. Soc. Lond. Ser. B-Biol. Sci.* **357**, 1625–1642 (2002).

66. Y. Xu, B. Labedan, N. Glansdorff, Surprising Arginine Biosynthesis: a Reappraisal of the Enzymology and Evolution of the Pathway in Microorganisms. *Microbiol. Mol. Biol. Rev.* **71**, 36–47 (2007).

67. E. N. Trifonov, Consensus temporal order of amino acids and evolution of the triplet code. *Gene* **261**, 139–151 (2000).

68. T. Janas, J. J. Widmann, R. Knight, M. Yarus, Simple, recurring RNA binding sites for L-arginine. *Rna* **16**, 805–16 (2010).

69. R. Amirnovin, An Analysis of the Metabolic Theory of the Origin of the Genetic Code. *J Mol Evol* **44**, 473–476 (1997).

70. T. A. Ronneberg, L. F. Landweber, S. J. Freeland, Testing a biosynthetic theory of the genetic code: fact or artifact? *Proc Natl Acad Sci USA* **97**, 13690–5. (2000).

71. S. E. Massey, Genetic Code Error Minimization as a Non-Adaptive But Beneficial Trait. *J. Mol. Evol.* **87**, 4–6 (2019).

72. R. D. Knight, S. J. Freeland, L. F. Landweber, Selection, history and chemistry: the three faces of the genetic code. *Trends Biochem Sci* **24**, 241–7 (1999).

73. E. V. Koonin, A. S. Novozhilov, Origin and Evolution of the Universal Genetic Code. *Annu. Rev. Genet.* **51**, 45–62 (2017).

74. D. Leister, T. Kleine, “Chapter three - Role of Intercompartmental DNA Transfer in Producing Genetic Diversity” in *International Review of Cell and Molecular Biology*, K. W. Jeon, Ed. (Academic Press, 2011), pp. 73–114.

75. T. Gabaldón, M. A. Huynen, From endosymbiont to host-controlled organelle: the hijacking of mitochondrial protein synthesis and metabolism. *PLoS Comput. Biol.* **3**, e219 (2007).

76. N. V. Chumachenko, Y. Novikov, M. Yarus, Rapid and simple ribozymic aminoacylation using three conserved nucleotides. *J Am Chem Soc* **131**, 5257–63 (2009).

77. R. M. Turk, M. Illangasekare, M. Yarus, Catalyzed and spontaneous reactions on ribozyme ribose. *J Am Chem Soc* **133**, 6044–50 (2011).

78. M. Yarus, The meaning of a minuscule ribozyme. *Philos Trans R Soc Lond B Biol Sci* **366**, 2902–2909 (2011).

79. 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).

80. C. W. Carter, What RNA World? Why a Peptide/RNA Partnership Merits Renewed Experimental Attention. *Life Basel Switz.* **5**, 294–320 (2015).

81. S. Tagami, J. Attwater, P. Holliger, Simple peptides derived from the ribosomal core potentiate RNA polymerase ribozyme function. *Nat. Chem.* **9**, 325–332 (2017).

82. A. Khvorova, Y. G. Kwak, M. Tamkun, I. Majerfeld, M. Yarus, RNAs that bind and change the permeability of phospholipid membranes. *Proc Natl Acad Sci USA* **96**, 10649–54 (1999).

83. A. Vlassov, A. Khvorova, M. Yarus, Binding and disruption of phospholipid bilayers by supramolecular RNA complexes. *Proc Natl Acad Sci USA* **98**, 7706–11 (2001).

84. T. Janas, T. Janas, M. Yarus, Specific RNA binding to ordered phospholipid bilayers. *Nucleic Acids Res* **34**, 2128–36 (2006).

85. T. Czerniak, J. P. Saenz, Lipid membranes modulate the activity of RNA through sequence-dependent interactions. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2119235119 (2022).

86. S. F. Banani, H. O. Lee, A. A. Hyman, M. K. Rosen, Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017).

87. E. Schmitt, P.-D. Coureux, A. Monestier, E. Dubiez, Y. Mechulam, Start Codon Recognition in Eukaryotic and Archaeal Translation Initiation: A Common Structural Core. *Int. J. Mol. Sci.* **20**, 939 (2019).

88. A. M. Burroughs, L. Aravind, The Origin and Evolution of Release Factors: Implications for Translation Termination, Ribosome Rescue, and Quality Control Pathways. *Int. J. Mol. Sci.* **20**, 1–24 (2019).

89. J. M. Ogle, *et al.*, Recognition of Cognate Transfer RNA by the 30S Ribosomal Subunit. *Science* **292**, 897–902 (2001).

90. D. W. Schultz, M. Yarus, tRNA structure and ribosomal function. I. tRNA nucleotide 27-43 mutations enhance first position wobble. *J Mol Biol* **235**, 1381–94 (1994).

91. T. Dale, R. P. Fahlman, M. Olejniczak, O. C. Uhlenbeck, Specificity of the ribosomal A site for aminoacyl-tRNAs. *Nucleic Acids Res.* **37**, 1202–1210 (2009).

92. I. Agmon, A. Bashan, R. Zarivach, A. Yonath, Symmetry at the active site of the ribosome: structural and functional implications. *Biol. Chem.* **386**, 833–844 (2005).

93. J. C. Bowman, A. S. Petrov, M. Frenkel-Pinter, P. I. Penev, L. D. Williams, Root of the Tree: The Significance, Evolution, and Origins of the Ribosome. *Chem. Rev.* **120**, 4848–4878 (2020).

94. L. Ribas de Pouplana, The evolution of aminoacyl-tRNA synthetases: From dawn to LUCA. *The Enzymes* **48**, 11–37 (2020).

95. M. Yarus, Getting past the RNA world: the initial Darwinian ancestor. *Cold Spring Harb. Perspect. Biol.* **3** (2011).

96. K. Vetsigian, C. Woese, N. Goldenfeld, Collective evolution and the genetic code. *Proc Natl Acad Sci USA* **103**, 10696–701 (2006).

97. A. J. Crapitto, A. Campbell, A. Harris, A. D. Goldman, A consensus view of the proteome of the last universal common ancestor. *Ecol. Evol.* **12**, e8930 (2022).

98. W.-K. Mat, H. Xue, J. T.-F. Wong, The genomics of LUCA. *Front. Biosci. J. Virtual Libr.* **13**, 5605–5613 (2008).

Legends

Table 1 – Cognate codons and anticodons in RNA-amino acid binding sites. Description, references and details are found in (23). RNAs that bound to columns containing immobilized amino acids were isolated from randomized ribooucleotide sequences. Probabilities shown for abundance of coding triplets in amino acid binding sites were assessed by comparing originally randomized non-site RNA sequences with their associated binding site sequences, assuming that these abundances are equal.

Figure 1 - Coding table initiation alters tolerance for later assignments. Frequencies for near-complete codes (>= 20 assigned functions) with exact SGC assignments (mis0) and near-complete codes with one difference from SGC (mis1) versus the fraction of non-SGC (randomized, Prand) assignments throughout coding table evolution. The two pairs of curves compare evolution after code initiation with a single SGC triplet with initiation with evolution after a full complement of stereochemically-determined triplets (_resi #2; see supplementary information).

Figure 2 – Coding table initiation alters acceptance of discordant coding assignments. A complete, full coding table having no assignments in common with the SGC (ALT, see supplementary information) was incorporated into evolving SGC-like codes, offering patch sizes of 1 to 5 randomly-chosen ALT triplets at a time. The number of accepted ALT-like assignments is plotted versus the number of triplets potentially transferred together. Each panel compares mean ALT-like assignment in the most complete codes (>= 20 encoded functions) with the mean in all evolving environmental coding tables at the time when the complete code arose.

Figure 2A – Code evolution initiated with a single SGC triplet.

Figure 2B – Code evolution initiated with Table 1's experimental triplets only. See supplementary information for "Stereo" initiating assignments; 8 chemically-determined triplets.

Figure 2C – Code evolution initiated with a full chemically-determined triplet set , _resi2. See supplementary information for _resi2; 20 chemically-determined triplets.

Figure 3 – Code evolution is determined by the number of functions assigned by an initial coding set. The experimental triplet set (Table 1) surveys 8 functions (called 'Stereo'). In contrast, _resi1, _resi2, _resi3, _resi4 and _resi5 are extended stereochemical triplet sets appropriate for complete coding tables (each with 20 determined triplets; see text **Chemistry for a code** and supplementary information)**.**

Figure 3A – Incompatible coding assignments increase both time to evolve >= 20 functions and ALT triplet incorporation. All codes evolved to encode >= 20 functions.

Figure 3B – Incompatible coding assignment decreases, and fraction of exact SGC coding increases, with the number of functions supplied by chemically-determined initiation. All codes evolved to encode >= 20 functions.

Figure 4A – **Unambiguous codes among fusion survivors converge to a consensus.** Code fusions that assign multiple meanings to individual triplets are less fit. Their extinction increases as codes become more complete, leaving more homogeneous, consensus codes for further evolution.

Figure 4B – **An RNA era for the code.** As the only evident means of specific code initiation, chemical interactions between oligoribonucleotides and free amino acids define an early RNA coding era (see **RNA Coding in context**). Such coding is sufficiently complete to produce the RNP transition, yielding highly competent structural and catalytic peptides. Stereochemical assignments tend to exclude independently-originated coding assignments, but these may be added to stereochemical precursor codes during a later endosymbiotic period.

Methods

Monte Carlo kinetic calculations (1, 45) were performed in console mode within the Pascal development environment Lazarus 2.2.ORC1 using the FPC compiler 3.2.2, hosted on a Dell computer running 64-bit windows 10 Enterprise, v. 22H2, with an Intel i9-8950HK CPU @ 2.90GHz and 32 GB of RAM.

Figure 1 results - Effect of differing initiations on later code acceptability.

1500 evolutionary environments ultimately containing ≈ 63,000 to 72,000 total coding tables were surveyed after initiation with a single SGC codon, and evolved until a code with >= 20 assigned functions appeared in the environment. Assignments after the first had probability Prand of being random, thus were not constrained to be SGC-like. Pmut=0.00975, Pdec=0.00975, Pinit=0.150, Pfus=0.00100, Ptab=0.080, Pwob=0.0000, Pdivn=0.500. Pmut is the probability/passage of a related assignment to an unocupied neighbor codon related by a single nucleotide change, Pdec is the probability/passage of selected codon assignment decay, Pinit is the probability/passage of initial assignment to a new coding table, Pfus is the (probability/passage-coexisting table) that randomly chosen codes will fuse, Ptab is the probability/passage of initiation of a new coding table, Pwob is the probability/passage of the initiation of wobble coding (always 0.0, late coding, here), Pdivn is the probability/passage that a random existing coding table will divide identically (1, 45).

1500 environments contained ≈ 35,000 to 48,000 tables initiated with full stereochemical _resi2 (see supplementary information), then evolved until a code with >= 20 assigned functions appeared in the environment. Assignments after the initial set had probability Prand of being random, thus were not constrained to be SGC-like.

Figure 2 results - Differing initiations restrict acceptance of alternative assignments.

Figure 2A – Codes Initiate evolution with one SGC codon - 1500 evolutionary environments ultimately containing ≈ 96,000 to 109,000 total coding tables were evolved to yield a code with >= 20 assigned functions, as in Figure 1, except that with probability Ptab per passage, a group of 1 to 5 assignments could come, randomly, from coding table ALT (supplementary information) which has no assignments in common with the SGC. Pmut=0.00975, Pdec=0.00975, Pinit=0.150, Pfus=0.00100, Ptab=0.040, Pwob=0.0000, Pdivn=0.500. In addition, evolution included random assignments at Prand = 0.15.

Figure 2B – Codes Initiate evolution with 8 experimental stereochemical triplets only (Stereo; Table 1) - 1500 evolutionary environments ultimately containing ≈ 69,000 to 90,000 total coding tables were evolved to yield a code with >= 20 assigned functions, as in Figure 2A, except that with probability Ptab per passage, a group of 1 to 5 assignments could come, randomly, from coding table ALT (supplementary information) which has no assignments in common with the SGC. In addition, evolution included random assignments at Prand = 0.15.

Figure 2C – Codes Initiate evolution with 20 stereochemical triplets (see text: **Chemistry for a code)** - 1500 evolutionary environments ultimately containing ≈ 32,000 to 37,000 total coding tables were initiated with _resi2 and evolved to yield a code with >= 20 assigned functions, as in Figure 2A, except that with probability Ptab per passage, a group of 1 to 5 assignments could come, randomly, from coding table ALT (supplementary information) which has no assignments in common with the SGC. In addition, evolution included random assignments at Prand = 0.15.

Figure 3 results - Complete stereochemical initiations differ markedly in final code properties

Completed stereochemical initiations _resi1 through _resi5 (20 stereochemically-determined triplets: described in **Chemistry for a code**) are shown in supplementary data. Probabilities are as in Fig. 2A.

Figure 3A – Time to evolve and ALT assignments for different initiations are related. Coding tables evolved to >= 20 assignments with possible ALT assignments and initiation with experimental stereochemical assignments from Table 1 ('Stereo') or estimated full-coding-table assignments _resi1, _resi2, _resi3, _resi4 and _resi5 (text **Chemistry for a code**; see also supplementary information). Other conditions as in Figure 2A.

Figure 3B – Stereochemical initiations differ because they differ in number of initially encoded functions. ALT assignments accepted and the fraction of very SGC-like codes (>= 20 assigned functions and entirely SGC assignments) are plotted versus number of stereochemical functions initially encoded. Other conditions as in Figure 2A.

probability	3.9×10^{-5}							
	UCG							
probability	4.0×10^{-26}		1.6×10^{-8}	3.2×10^{-131}		5.5×10^{-5}	2.7×10^{-13}	6.0×10^{-6}
anticodons	**CCU**		**GUG**	**UAU**		**GAA**	**CCA**	**AUA**
amino acids	***Arg***	***Gln***	***His***	***Ile***	***Leu***	***Phe***	***Trp***	***Tyr***
codons	**AGG**			**AUU**				
probability	5.6×10^{-19}			8.0×10^{-110}				

Table 1

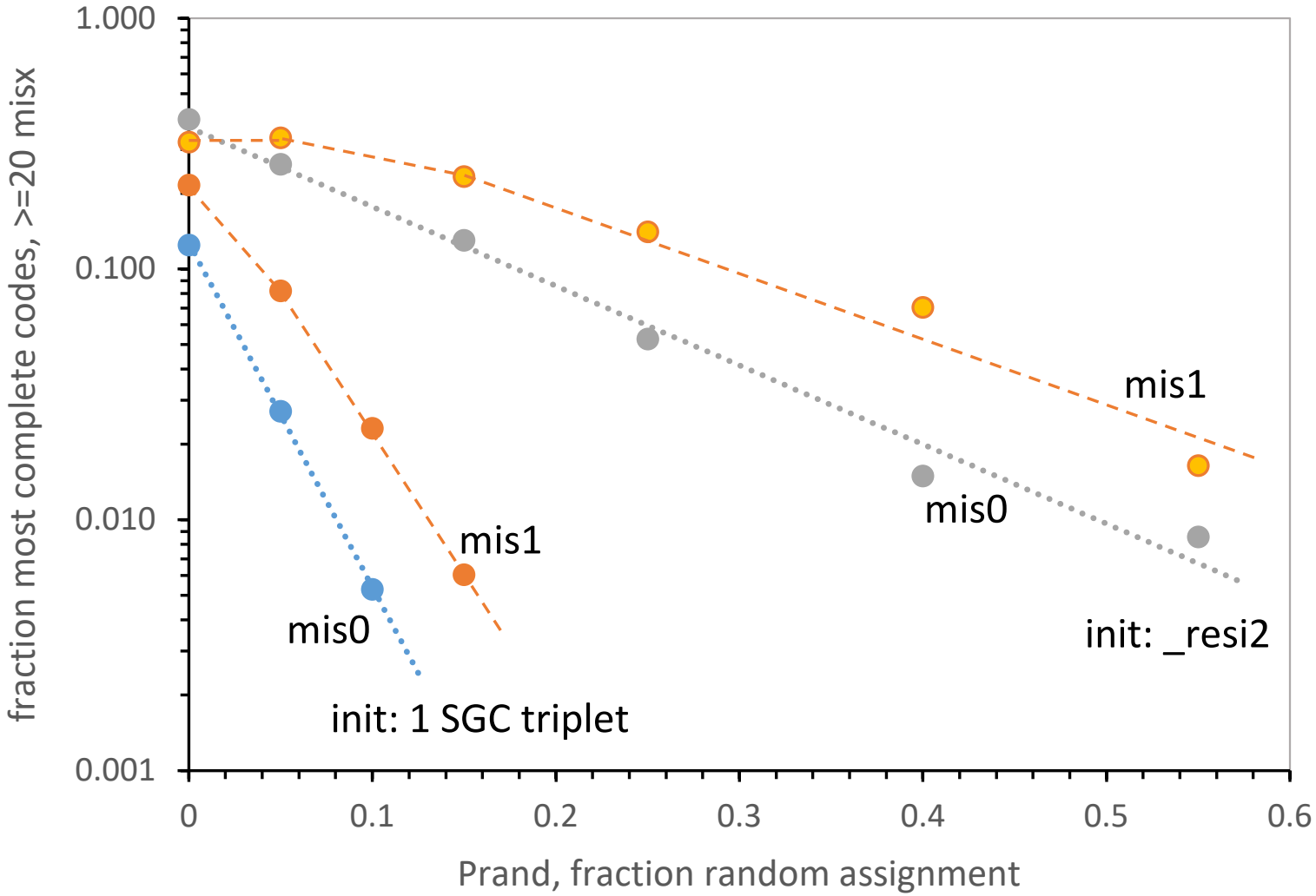
1.000
0.100
0.010
0.001
fraction most complete codes, >=20 misx
0
0.1
0.2
0.3
0.4
0.5
0.6
Prand, fraction random assignment
mis1
mis0
mis1
mis0
init: _resi2
init: 1 SGC triplet

Figure 1

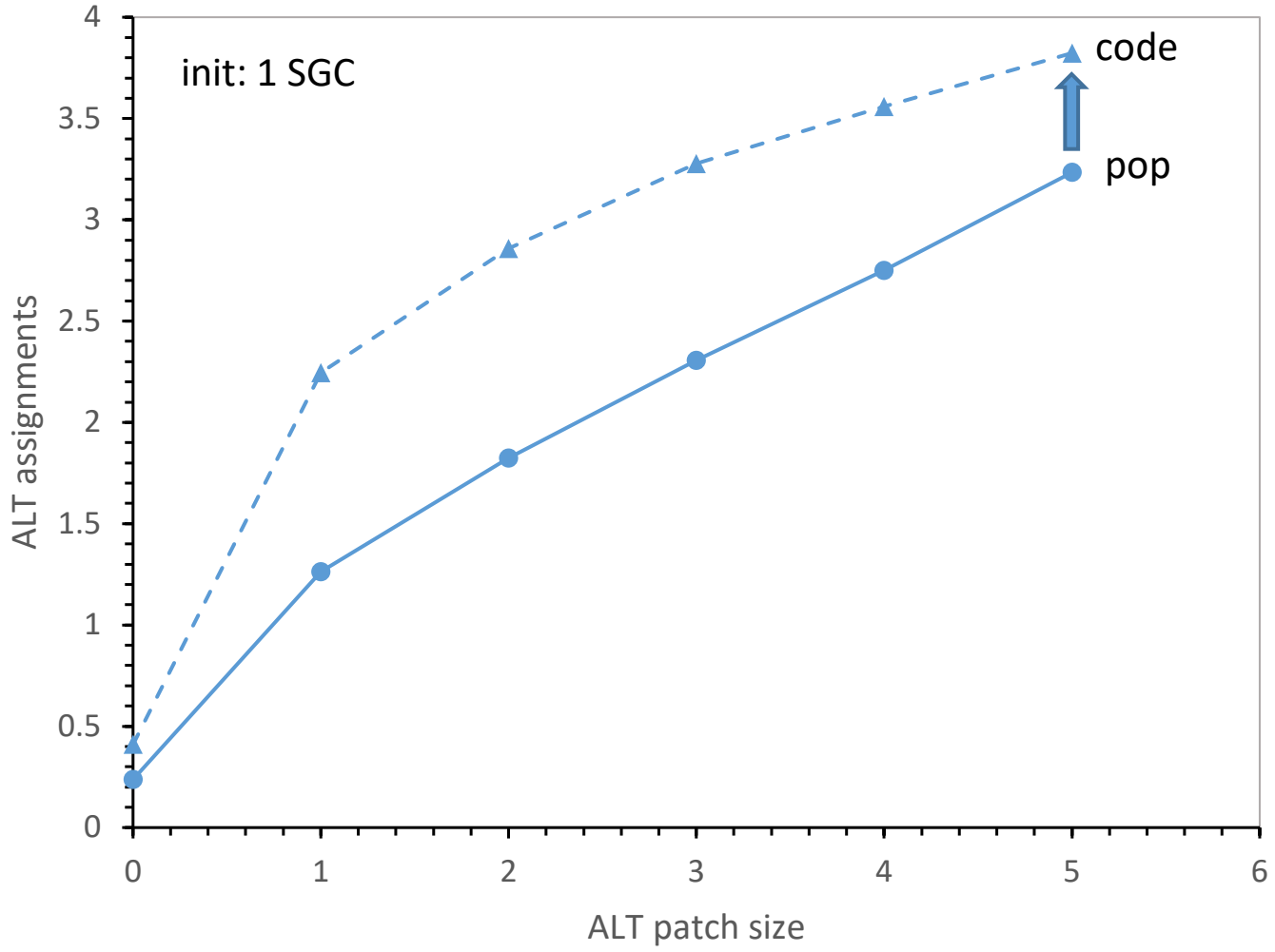
init: 1 SGC
code
pop
ALT assignments
ALT patch size
0
0.5
1
1.5
2
2.5
3
3.5
4
0
1
2
3
4
5
6

Figure 2A

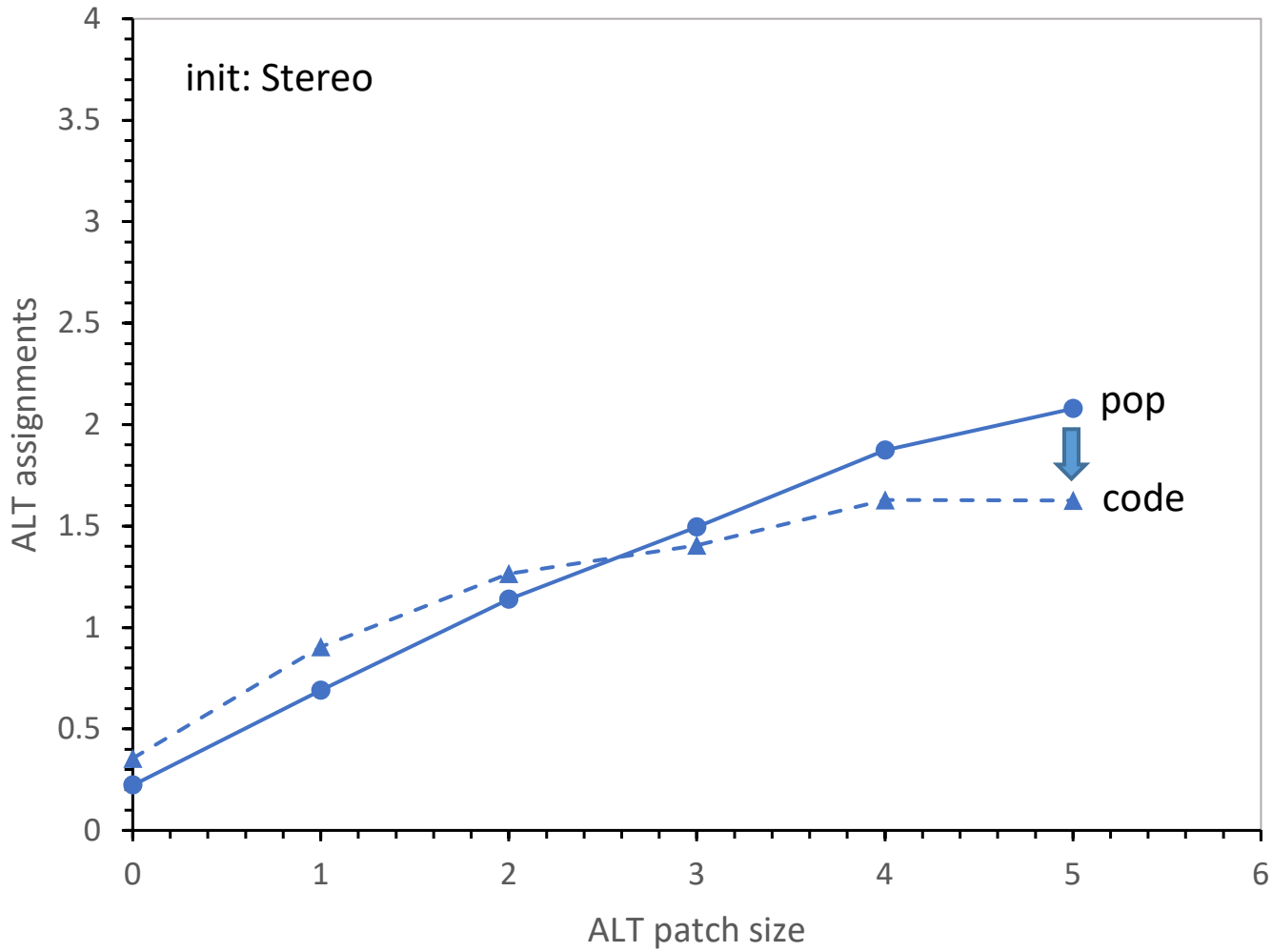
init: Stereo
pop
code
ALT assignments
ALT patch size
0
0.5
1
1.5
2
2.5
3
3.5
4
0
1
2
3
4
5
6

Figure 2B

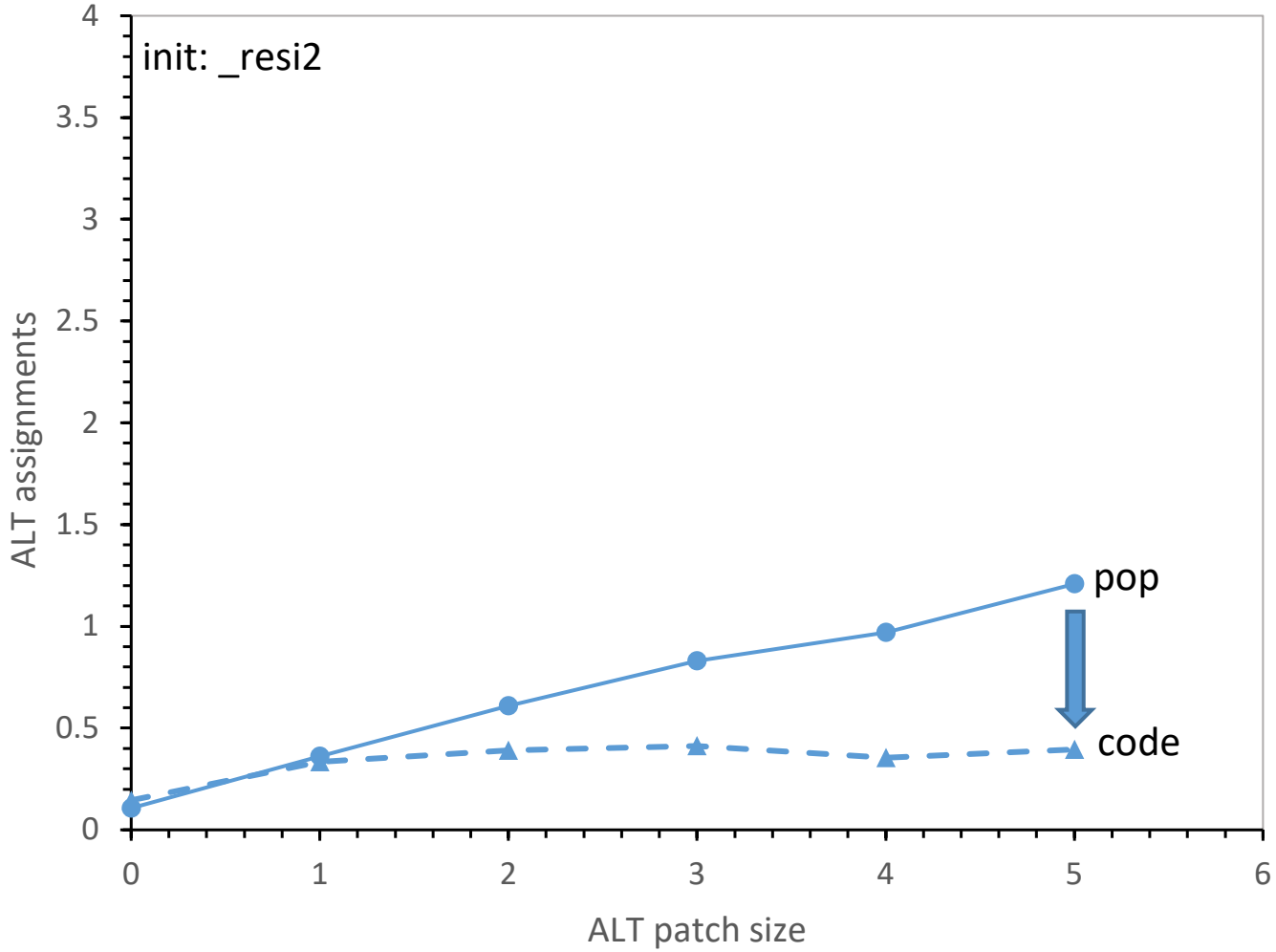
init: _resi2
ALT assignments
ALT patch size
pop
code
0
0.5
1
1.5
2
2.5
3
3.5
4
0
1
2
3
4
5
6

Figure 2C

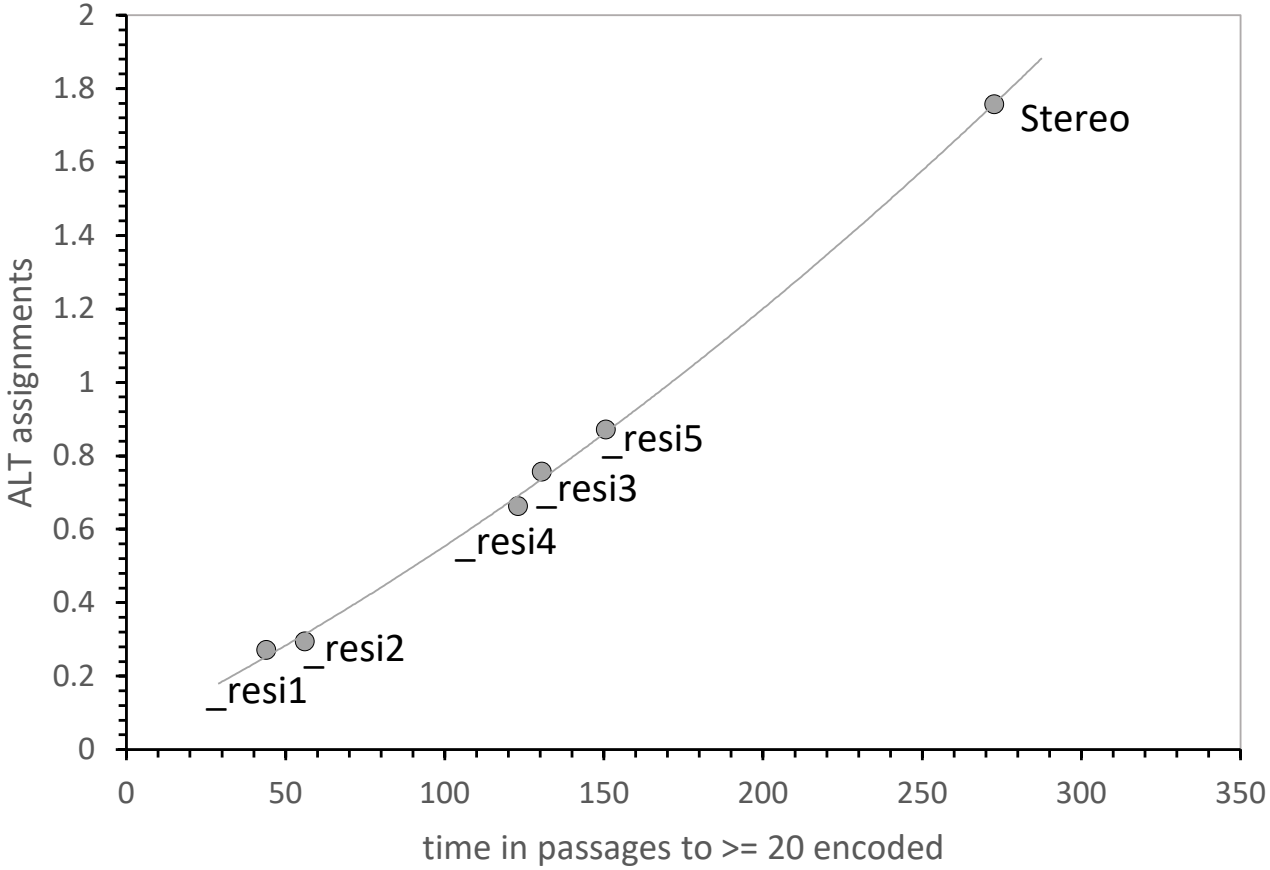
ALT assignments
time in passages to >= 20 encoded
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
0
50
100
150
200
250
300
350
Stereo
_resi5
_resi3
_resi4
_resi2
_resi1

Figure 3A

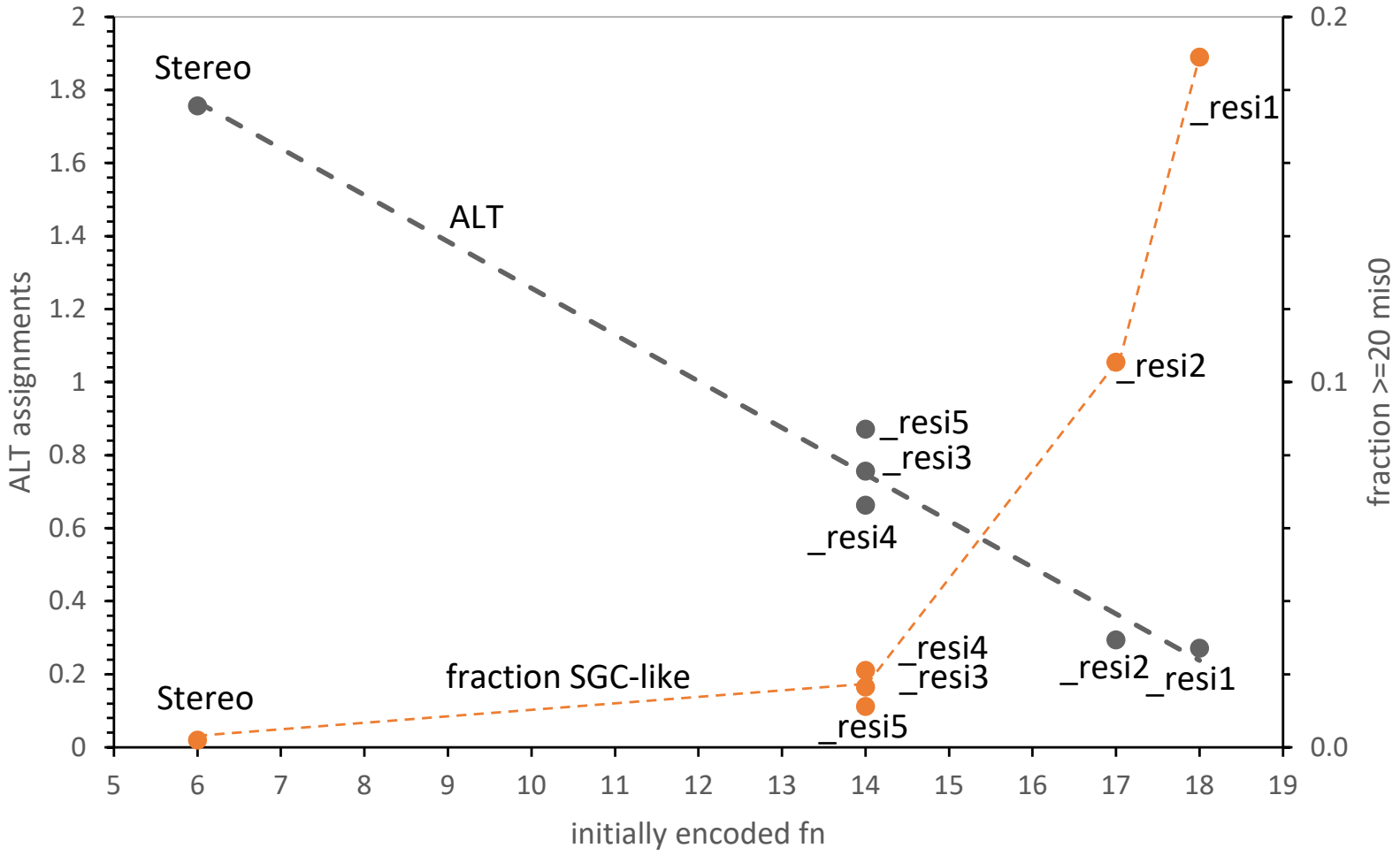
ALT assignments
fraction >=20 mis0
initially encoded fn
Stereo
ALT
_resi1
_resi2
_resi5
_resi3
_resi4
fraction SGC-like
Stereo
_resi4
_resi3
_resi5
_resi2
_resi1
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0.2
0.1
0.0
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

Figure 3B

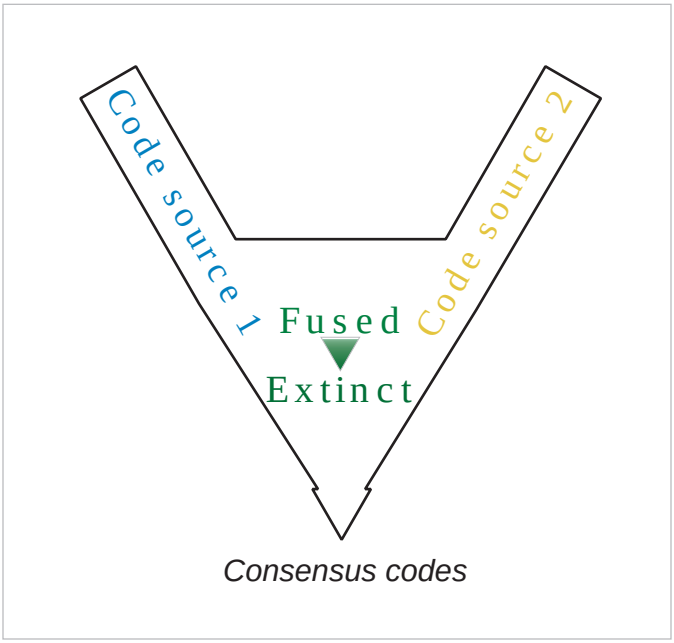

Figure 4A

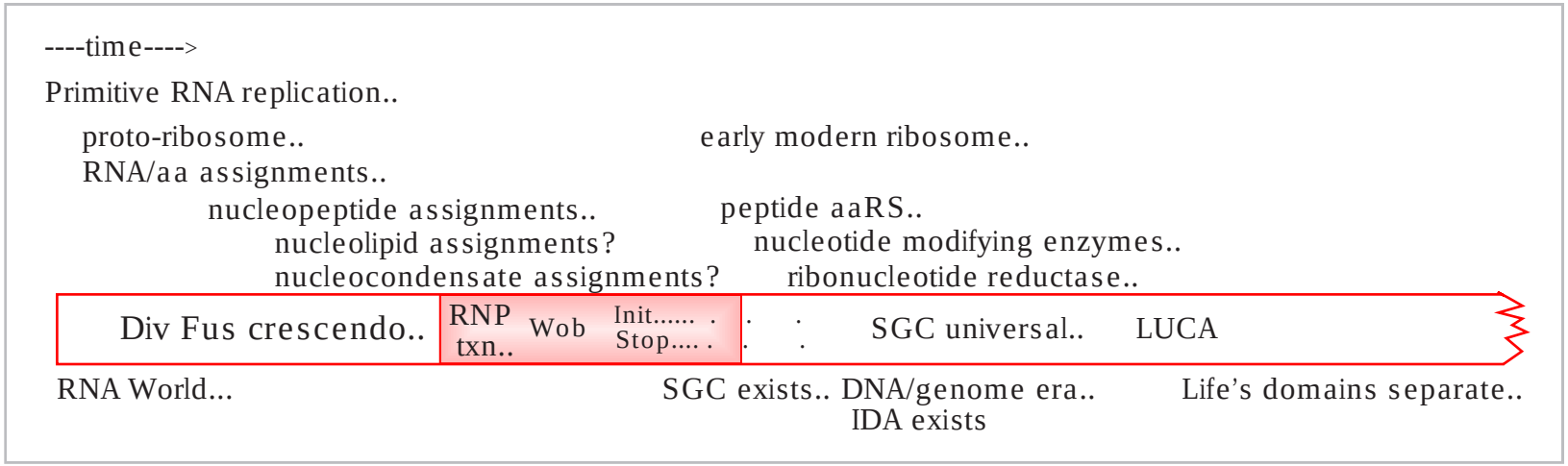
----time---->
Primitive RNA replication..
proto-ribosome..
early modern ribosome..
RNA/aa assignments..
nucleopeptide assignments..
peptide aaRS..
nucleolipid assignments?
nucleotide modifying enzymes..
nucleocondensate assignments?
ribonucleotide reductase..
Div Fus crescendo..
RNP txn..
Wob
Init..... Stop....
SGC universal..
LUCA
RNA World...
SGC exists..
DNA/genome era..
IDA exists
Life's domains separate..

Figure 4B

	U	C	A	G	
	Glu(2)	Leu(6)	Leu(6)	Arg(6)	U
U	Glu(2)	Leu(6)	Leu(6)	Arg(6)	C
	Tyr(2)	Leu(6)	His(2)	Trp(1)	A
	Tyr(2)	Leu(6)	His(2)	STP(3)	G
	Ser(6)	Gly(4)	STP(3)	Val(4)	
C	Ser(6)	Gly(4)	STP(3)	Val(4)	
	Ser(6)	Gly(4)	Lys(2)	Val(4)	
	Ser(6)	Gly(4)	Lys(2)	Val(4)	
	Ala(4)	Arg(6)	Gln(2)	Asp(2)	
A	Ala(4)	Arg(6)	Gln(2)	Asp(2)	
	Ala(4)	Arg(6)	Asn(2)	Pro(5)	
	Ala(4)	Arg(6)	Asn(2)	Cys(1)	
	Thr(4)	Ile(3)	Ser(6)	Pro(5)	
G	Thr(4)	Ile(3)	Ser(6)	Pro(5)	
	Thr(4)	Ile(3)	Phe(2)	Pro(5)	
	Thr(4)	Met(1)	Phe(2)	Pro(5)	

Supplementary information: coding tables

ALT
full,
complete,
alternative
code

(n) = n assignments

SGC

	U	C	A	G	
	Phe(2)	Ser(6)	Tyr(2)	Cys(2)	U
U	Phe(2)	Ser(6)	Tyr(2)	Cys(2)	C
	Leu(6)	Ser(6)	STP(3)	STP(3)	A
	Leu(6)	Ser(6)	STP(3)	Trp(1)	G
	Leu(6)	Pro(4)	His(2)	Arg(6)	
C	Leu(6)	Pro(4)	His(2)	Arg(6)	
	Leu(6)	Pro(4)	Gln(2)	Arg(6)	
	Leu(6)	Pro(4)	Gln(2)	Arg(6)	
	Ile(3)	Thr(4)	Asn(2)	Ser(6)	
A	Ile(3)	Thr(4)	Asn(2)	Ser(6)	
	Ile(3)	Thr(4)	Lys(2)	Arg(6)	
	Met(1)	Thr(4)	Lys(2)	Arg(6)	
	Val(4)	Ala(4)	Asp(2)	Gly(4)	
G	Val(4)	Ala(4)	Asp(2)	Gly(4)	
	Val(4)	Ala(4)	Glu(2)	Gly(4)	
	Val(4)	Ala(4)	Glu(2)	Gly(4)	

Stereo

	U	C	A	G	
	zip(56)	zip(56)	Tyr(1)	zip(56)	U
U	Phe(1)	zip(56)	zip(56)	zip(56)	C
	zip(56)	zip(56)	zip(56)	zip(56)	A
	zip(56)	zip(56)	zip(56)	Trp(1)	G
	zip(56)	zip(56)	zip(56)	zip(56)	
C	zip(56)	zip(56)	His(1)	zip(56)	
	zip(56)	zip(56)	zip(56)	Arg(2)	
	zip(56)	zip(56)	zip(56)	zip(56)	
	Ile(2)	zip(56)	zip(56)	zip(56)	
A	zip(56)	zip(56)	zip(56)	zip(56)	
	Ile(2)	zip(56)	zip(56)	zip(56)	
	zip(56)	zip(56)	zip(56)	Arg(2)	
	zip(56)	zip(56)	zip(56)	zip(56)	
G	zip(56)	zip(56)	zip(56)	zip(56)	
	zip(56)	zip(56)	zip(56)	zip(56)	
	zip(56)	zip(56)	zip(56)	zip(56)	

zip = no assignment

_resi1

	U	C	A	G	
	zip(44)	zip(44)	Tyr(1)	Cys(1)	U
U	Phe(1)	zip(44)	zip(44)	zip(44)	C
	zip(44)	zip(44)	STP(1)	zip(44)	A
	zip(44)	zip(44)	zip(44)	Trp(1)	G
	zip(44)	zip(44)	zip(44)	zip(44)	
C	zip(44)	zip(44)	His(1)	zip(44)	
	zip(44)	Pro(1)	zip(44)	Arg(2)	
	zip(44)	zip(44)	zip(44)	zip(44)	
	Ile(2)	Thr(1)	Asn(1)	zip(44)	
A	zip(44)	zip(44)	zip(44)	zip(44)	
	Ile(2)	zip(44)	zip(44)	zip(44)	
	Met(1)	zip(44)	Lys(1)	Arg(2)	
	zip(44)	zip(44)	zip(44)	zip(44)	
G	Val(2)	Ala(1)	zip(44)	zip(44)	
	Val(2)	zip(44)	zip(44)	zip(44)	
	zip(44)	zip(44)	Glu(1)	Gly(1)	

zip = no assignment

_resi2

	U	C	A	G	
	zip(44)	Ser(2)	Tyr(1)	zip(44)	U
U	Phe(1)	zip(44)	zip(44)	zip(44)	C
	zip(44)	Ser(2)	STP(1)	zip(44)	A
	zip(44)	zip(44)	zip(44)	Trp(1)	G
	zip(44)	zip(44)	zip(44)	zip(44)	
C	zip(44)	zip(44)	His(1)	zip(44)	
	zip(44)	zip(44)	zip(44)	Arg(2)	
	zip(44)	Pro(1)	Gln(1)	zip(44)	
	Ile(2)	zip(44)	zip(44)	zip(44)	
A	zip(44)	Thr(1)	Asn(1)	zip(44)	
	Ile(2)	zip(44)	zip(44)	zip(44)	
	Met(1)	zip(44)	zip(44)	Arg(2)	
	zip(44)	zip(44)	zip(44)	zip(44)	
G	zip(44)	zip(44)	Asp(1)	zip(44)	
	zip(44)	Ala(1)	Glu(2)	zip(44)	
	zip(44)	zip(44)	Glu(2)	zip(44)	

zip = no assignment

_resi3

	U	C	A	G	
	zip(44)	Ser(1)	Tyr(1)	zip(44)	U
U	Phe(1)	zip(44)	zip(44)	zip(44)	C
	zip(44)	zip(44)	zip(44)	zip(44)	A
	zip(44)	zip(44)	STP(1)	Trp(1)	G
	zip(44)	zip(44)	His(2)	Arg(4)	
C	zip(44)	Pro(2)	His(2)	Arg(4)	
	zip(44)	zip(44)	zip(44)	Arg(4)	
	zip(44)	Pro(2)	zip(44)	zip(44)	
	Ile(2)	Thr(1)	zip(44)	zip(44)	
A	zip(44)	zip(44)	zip(44)	zip(44)	
	Ile(2)	zip(44)	zip(44)	zip(44)	
	Met(1)	zip(44)	Lys(1)	Arg(4)	
	zip(44)	zip(44)	zip(44)	zip(44)	
G	Val(2)	zip(44)	zip(44)	zip(44)	
	Val(2)	zip(44)	zip(44)	zip(44)	
	zip(44)	zip(44)	zip(44)	zip(44)	

zip = no assignment

_resi4

	U	C	A	G	
	zip(44)	zip(44)	Tyr(1)	Cys(1)	U
U	Phe(1)	zip(44)	zip(44)	zip(44)	C
	zip(44)	zip(44)	STP(2)	zip(44)	A
	zip(44)	Ser(1)	STP(2)	Trp(1)	G
	zip(44)	zip(44)	zip(44)	zip(44)	
C	zip(44)	zip(44)	His(1)	zip(44)	
	zip(44)	zip(44)	zip(44)	Arg(2)	
	zip(44)	zip(44)	zip(44)	zip(44)	
	Ile(2)	zip(44)	zip(44)	zip(44)	
A	zip(44)	zip(44)	zip(44)	zip(44)	
	Ile(2)	zip(44)	Lys(1)	zip(44)	
	zip(44)	Thr(1)	zip(44)	Arg(2)	
	Val(3)	zip(44)	zip(44)	zip(44)	
G	Val(3)	Ala(2)	zip(44)	zip(44)	
	zip(44)	Ala(2)	zip(44)	zip(44)	
	Val(3)	zip(44)	Glu(1)	zip(44)	

zip = no assignment

_resi5

	U	C	A	G	
	zip(44)	Ser(4)	Tyr(1)	zip(44)	U
U	Phe(1)	zip(44)	zip(44)	Cys(1)	C
	zip(44)	zip(44)	STP(1)	zip(44)	A
	zip(44)	Ser(4)	zip(44)	Trp(1)	G
	zip(44)	Pro(3)	zip(44)	zip(44)	
C	zip(44)	Pro(3)	His(1)	zip(44)	
	zip(44)	zip(44)	zip(44)	Arg(2)	
	zip(44)	Pro(3)	zip(44)	zip(44)	
	Ile(2)	zip(44)	zip(44)	Ser(4)	
A	zip(44)	zip(44)	zip(44)	Ser(4)	
	Ile(2)	zip(44)	zip(44)	zip(44)	
	zip(44)	zip(44)	zip(44)	Arg(2)	
	zip(44)	zip(44)	zip(44)	zip(44)	
G	zip(44)	Ala(1)	zip(44)	zip(44)	
	zip(44)	zip(44)	Glu(1)	zip(44)	
	zip(44)	zip(44)	zip(44)	Gly(1)	

zip = no assignment