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, division, frozen accident, adaptation, error minimization, stereochemistry, Lateral Gene Transfer, Horizontal Gene Transfer

Abstract

Multiple prior experiments show that RNA binds chemically varied amino acids within specific ribo-oligonucleotide sequences. The smallest, simplest, and therefore most likely primitive RNA binding sites contain functional triplets corresponding to the Standard Genetic Code (SGC). Here, the implications of such observed coding triplets are calculated, combining them with an optimized kinetic model for SGC evolution. RNA-amino acid interactions at known frequencies both choose an SGC-like code, and by the same mechanism, effectively resist alternative triplet assignments. Resistance to external coding is evident in varied code initiation scenarios. Observed RNA-mediated assignments are also likely sufficient to produce the "ribonucleopeotide (RNP) transition" to a modern RNP code. This can fully account for extreme selection of the SGC among its astronomical code possibilities; very SGC-like codes are ≈ 1/50 to 1/5 of codes within such a population. Nevertheless, complete accounting will depend on RNA affinities still unknown. Such a code begins as mostly stereochemical, excludes other assignments, and critically relies on properties unique to fusible microbes. After its RNP transition, other assignment mechanisms (adaptation, co-evolution, revised stereochemistry, LGT) likely expand the code because cellular intermediates with ≥ 1 code exist, allowing time for mutual code exchanges. The 83 order-of-magnitude focus required to find an SGC was therefore spanned by at least two assignment mechanisms, in at least two successive eras.

Introduction

An extraordinary bottleneck. Allowing unassigned functions, there are 1.8 x 10^{83} possible genetic codes with 20 functions and 64 triplets (Yarus 2021b). Moreover, available primordial amino acids (Miller 1953) greatly exceed selected biological forms (Cleaves 2010). Carbonaceous meteorites contain at least 75 amino acids, a minority of which are biological (Pizzarello 2006). Further, it is clear that multiple bacterial codon assignments can be changed without severe biological consequences (Ostrov et al. 2016). Therefore, it is most striking that more than 250,000 codes in current bacterial and archaeal

genomes (Shulgina and Eddy 2021) present only variants of the Standard Genetic Code using the same 22 amino acids (Yuan et al. 2010).

All modern biota thus are descendants of one ancestral group whose translation was guided by a close relative of the SGC. 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 ancestral group. For example, consider the particular choice of encoded amino acids (Weber and Miller 1981; Cleaves 2010); some standard amino acids are probably conditional choices inherited from our universal coding ancestor. More generally: the SGC contains information about a crucial bottleneck in life's evolution.

Initial coding to the Standard Genetic Code. It is inevitable that numerous amino acids were assigned to coding sequences 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 active site geometries supporting catalyzed regiospecific aminoacyl transfer (Rubio Gomez and Ibba 2020). Thus the aaRS era must have arisen from a simpler type of competent aminoacyl-RNA synthesis and translation, presumably mediated by RNAs (Gilbert 1986; Illangasekare and Yarus 2012; Wilson and Lilley 2021). In particular, both RNA-catalyzed amino acid activation (Kumar and Yarus 2001) and activated aminoacyl transfer to RNA (e.g., Illangasekare and Yarus 2012) have been experimentally reproduced.

RNA-amino acid interactions. There is extensive evidence that amino acids bind specifically to small ribo-oligonucleotide sequences. Though originating with a natural *Tetrahymena* example, abundant other data have accumulated (Yarus 1988; Yarus and Christian 1989; Johnson and Wang 2010; Rodin et al. 2011; Sengupta et al. 2016; Kapral et al. 2022; Marton Menendez and Nesbitt 2022).

RNA binding sites for amino acids have also been isolated and characterized experimentally. Randomized RNAs were subjected to selection-amplification (Irvine et al. 1991) after affinity chromatography for amino acids (Ciesiolka et al. 1996). Elution with free amino acids ensures that selected RNAs bind free amino acids. 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 (Yarus 2017). Local sequences within such binding sites contain an excess of cognate coding triplets for their amino acids, when compared to nearby non-binding sequences, which came from the same randomized RNA pools and passed through an identical purification. Interestingly, if the smallest binding sequences (therefore the most likely primitive examples) are selected, inclusion of essential coding triplets is complete (Lozupone et al. 2003).

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 requires conserved ribonucleotide contacts. 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 (e.g., Majerfeld and Yarus 2005).

These sites are summarized in Table 1, which links amino acids, their coding triplets and the probability of observed triplet concentrations in RNA binding site nucleotides. Probability (that cognate triplets are equally distributed between site and non-site sequences) is calculated by comparison to initially randomized non-site nucleotides adjacent to selected binding sites. These nearby control sequences are covalently linked to site nucleotides, but are less or not at all essential to amino acid binding function. A low probability ceiling corrects for multiple trials (Yarus 2017): in Tab. 1, significant triplet concentrations have probabilities $< 2.3 \times 10^{-4}$.

This analysis of RNA-amino acid binding sites has been questioned (Koonin and Novozhilov 2008), but common objections to stereochemical mechanisms are not compelling. RNA-amino acid association does not depend on calculated site probabilities. Site populations consist of highly repeated independent examples of one, or more, families of related site sequences (one example: (Illangasekare and Yarus 2002). Thus, close association of cognate triplet sequences and amino acid binding sites is directly observed, and cannot be an artifact of statistical methods. In addition, these data do not identify coding triplets as contacts for the amino acids (an early idea tracing to coding in 'holes' on DNA: Gamow 1954). But this is not logically required, because early 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 generally mandates moderate dissociation constants (Yarus 2017), but this is not a valid objection to their importance. Michaelis-Menten analysis dictates that catalytic reaction rates depend on forward 1st order rates, as well as affinities. Millimolar or substantial micromolar association constants (Yarus 2017) are entirely consistent with rapid catalytic (bio)chemical reactions.

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

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

Variety. A variety of side chain chemistries exhibit RNA stereochemical links (Tab. 1): a potential hydrophobic core for folded structures (Ile), potentially catalytic (His), polar, strong H-bonding (Arg), planar aromatic (Phe, Tyr, Trp).

Specificity. For example, RNA affinities distinguish similar hydrophobic sidechains, isoleucine and leucine (Tab. 1). An L-valine site distinguishes D-valine, L-Leucine, L-isoleucine, and L-norvaline, chemically very similar (Majerfeld and Yarus 1994). An RNA binding site selected for phenylalanine can either accept other planar aromatics, or reject tyrosine and tryptophan (Illangasekare and Yarus 2002).

Triplets. No RNA affinities are associated with amino acid codons alone. Ile and Arg linked codons occur with linked anticodons to the same amino acid. Arg alone is linked to potentially paired, complementary anticodon/codons, as well as to a second, minimally-related anticodon. All six amino acids linked to stereochemistry therefore are bound in close proximity to their anticodons,

thereby resembling anticodon-cognate amino acid associations in modern translation (Holley et al. 1966; Crick 1968).

RNA-amino acid interactions provide a unique code origin. Above, the astronomically vast universe of possible codes is emphasized. Concentration on one of these possibilities must be explained. Ideally, this would explain choice of the SGC, but there is a similar problem for any other unique code. Selection of one of $\approx 10^{83}$ options will necessarily be non-trivial.

If we wish to associate RNA triplets and amino acids, an effective, specific mechanism for bringing RNA sequences together with amino acids is required. But a mechanism for this extraordinary focus almost identifies itself. Chemical affinity of amino acids for RNA structures containing essential future coding triplets meets this critical need.

Thus a stereochemical origin is unique, for two reasons. First, because of the need for choice among an astronomical code variety. Second, to juxtapose reactants and so catalyze encoded synthetic reactions. It is difficult to imagine encoded chemistry without a physical interaction inducing proximity between specific amino acids and coding sequences.

Frances Crick supposed that the SGC might be a 'frozen accident', and "in its extreme form", coding assignments could all be due to chance (Crick 1968). But the SGC is not random (Yarus 2017). Moreover, once evolved, the code would not be changed, because of the necessity to preserve pre-existing essential gene products (Sella and Ardell 2006). However, Crick allowed that some amino acids might be initially encoded stereochemically (Crick 1968). The finding that six of eight amino acids, of varied types, associate with cognate triplets (Tab. 1) suggests that chemical determinism rather than undirected chance has a major 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 (Freeland and Hurst 1998). Adaptive code evolution implies a pre-existing nascent code whose assignments will be guided toward minimal error. So, adaptation requires an independent way to initiate coding, and especially to 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 (Wong 1975; Di Giulio 2008). Coevolution is a persuasive idea, but offers no way to initiate coding, or to make its astronomically specific choices.

Given that the only apparent way to initiate coding is with specific stereochemical interaction (Yarus 1998; Ma 2010), below we seek an informative relation between chemical and other code-ordering mechanisms.

Chemistry for a code. Presumably cognate codon, anticodon, or both in a binding site can be counted as stereochemical links, useable in forming a later code based on binding (Yarus et al. 2005; Yarus 2017). 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 (Yarus 2017). In Tab. 1, 6 of 8 amino acids still show a stereochemical link if only anticodons are taken as significant (Weber and Lacey Jr 1978).

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 refer to these data concisely. While extrapolation is not unique, it is assumed below that 20 amino acids would implicate (20/8) x 8 = 20 stereochemically related triplet assignments for 20 SGC amino acids. Thus, on simple assumptions, ≈ (8/24) ≈ (20/61) ≈ 1/3 of amino acid-encoding triplets in the near-universal genetic code arguably arise 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.

A mechanism for coding table evolution. The path of least selection provides the most probable evolutionary route for a favored characteristic (Yarus 2022b). Least selection for genetic encoding utilizes fusion of separately evolving partial codes (Yarus 2021c). Such code fragments might have been selected to serve varied functions, so fusing codes would support continually increasing biological abilities. In addition, code fusions implement code evolution in parallel, so yielding complete codes more quickly (Yarus 2021c). Given a prominent role for code fusions (Yarus 2022a), either homogeneous or heterogeneous fusions leading to the SGC appear possible (Yarus 2023c). That is, fusions may join further-evolved descendants of one original code, or join codes from independent lines of descent.

The point of this development. Below, above findings are combined: experimental amino acid-RNA associations are followed within the shortest, most accurate code evolution known, to study their joint effects.

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 inferred in **Chemistry for a code** above. Using different initiations reveals early effects on the ultimate codes. Initial codes are completed (evolve to >= 20 assigned functions) by adding single random SGC-like assignments with differing accuracies (having a probability Prand of random assignment and thus approximately, of differing from 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 and the lower curves for single-codon initial assignment.

Large differences appear from distinct beginnings; initiation of code evolution with observed levels of stereochemical assignment makes completion of the code much more tolerant to error (expressed as randomness among later assignments, Prand). Thus, later assignments are allowed to have more errors, while still allowing close evolutionary approach to the SGC. So, stereochemical initiation gathers other compatible assignments to its growing code, 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 moderate random (instead of SGC) coding (Fig. 1).

A**mino acid-RNA interactions reduce acceptability of divergent assignments.** Just above, greater early SGC-like stereochemical assignment favors later SGC-like assignment. There is a complementary effect on non-SGC assignments; early SGC-like assignments disfavor infusion of later, conflicting assignments.

Fig. 2 compares incorporation from an entirely different coding scheme into near-complete SGC codes, using three kinds of initial coding. Fig. 2A shows data for random initiation, 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 kind of initiation, incorporation of an alternative code that has no assignments in common with the SGC (ALT, in supplementary data) into the total population of codes (labeled **pop**) is compared to incorporation into the most complete code (labeled **code**).

All panels have the same Y-axis. This emphasizes that beginning with more stereochemistry, increasing from single SGC assignments, to Stereo, to _resi2, alternative coding decreases (Fig. 2A, 2B, 2C). Though this has a simple basis (increasing numbers of codons are committed early), it still shows a notable effect. For example, the mean rightward code in Fig. 2A has seen 76 alternative assignments 5 at a time, but accepted only 3.8 into an average SGC-like (>= 20 function mis 0) code.

Secondly, Fig. 2 distinguishes alternative incorporation into the entire code population (**pop**) from that into the most complete codes (**code**) for 1 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 (upward arrow relates population to most complete codes, Fig. 2A). Coding restricted to experimental amino acid-triplet initial assignments (Fig. 2B) is intermediate, actually crossing over at 3 triplets/alternative code. These intermediate initiations finally reject alien assignments, yielding most complete codes with fewer alternative codons (downward arrow) than the code population as a whole. Finally, initial assignment with stereochemistry calculated for total amino-acid-RNA sites ((Yarus 2017), Table1, Fig. 2C) discriminates strongly against alternative assignments, with fewer alternative encodings in most-complete codes than the code population at all 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 rejection of alternative encoding is particularly notable with full stereochemical initiations, which reject intrusive coding at every alternative code patch size, accepting a mean of < 1 foreign assignment.

Full stereochemical initiations differ greatly. For clarity, initiation with stereochemical assignments has been introduced above as if it were uniform. But it emphatically is not. Code extension _resi2 (supplementary data file) has been used for illustrations, but random extensions of known amino acid-RNA interactions have very 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. Points, as indicated, are for 8 triplets from evolved RNA binding sites (Stereo, Tab. 1, **Chemistry for a complete**

code), as well as for 5 different random extensions to 64 codons (supplementary data). In Fig. 3A, the five random extensions of experimental amino acid-RNA interactions have been named and are reproduced as standard coding tables in 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. When evolution of a stereochemically-initiated complete code requires more time, that code also admits proportionately more assignments from an independent coding system.

Varied stereochemical initiation efficiencies have a simple explanation. Differences between full stereochemical initiations are attributable to differing numbers of functions encoded. Time to evolve complete codes (Fig. 3A) and other evolutionary indices will vary with evolutionary distance to the SGC (Yarus 2022b). Because full initial encoding examples 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 observed 8 initial triplets differ in number of final functions, and therefore differ in initial evolutionary distance from the SGC.

This is reflected in Fig. 3B, where the number of admitted alternative 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 (Yarus 2023d) is plotted alongside time or alternative assignments. Such an inclusive distance measure reflects both amount of change and evolution 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 initial assignments (right ordinate, Fig. 3B). Observation of almost a fifth of near-complete codes identical to the SGC (Fig. 3B) concisely evokes the power of stereochemical initiation at known frequencies. 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 met via selection from code fusion-division-late-wobble evolution, initiated with a favorable stereochemical assignment set. Alternatively: given _resi1 initial assignments, 19% of such nascent codes span the 83 order-of-magnitude selection required for the SGC. Moreover, even for _resi3, _resi4 and _resi5, SGC evolution would need to select about 1 code in 50, a seemingly short gap and simple task.

Resisting inaccurate assignment. In Fig. 2 and Fig. 3, code assignments are 15% random (see legends). These evolving code populations therefore benefit from Fig. 1's effect. They find the SGC even if challenged by departure from a uniform set of assignment principles, 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 (Yarus 2021b); selecting a favorable, less probable minority among a large, less-fit population. In Methods, Fig. 2 populations range up to 10^5 individual codes. 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. Such microbes are also required to exploit the favored tail of a unimodally distributed code by truncation; that is, improving by accepting only rare best examples (Yarus 2022b).

Discussion

Optimized evolution of coding tables. We draw the implications of observed stereochemistry using a Monte Carlo kinetic model of code evolution (Yarus 2021b). The model has been optimized by determining the most rapid and accurate of 32 possible pathways for SGC evolution (Yarus 2023d). The calculation constructs codes assuming that code evolution obeys standard rules of chemical kinetics (Yarus 2021b). Different evolutionary pathways of can be compared using a principle of least selection (Yarus 2022b), which assumes that the code most rapidly evolved and adhering most closely to known assignments (of the Standard Genetic Code, SGC) will require least selection, that is, will be the most probable precursors of the historical SGC.

Selected major conclusions of prior analysis are that only broadly validated evolutionary means are required to reach the SGC (Yarus 2023d), that wobble was likely a later coding development, after most elementary assignments yarus (Yarus 2021b, 2023a), that such wobble was more likely to have been simplified Crick wobble (Yarus 2021a) than four-fold superwobble (Alkatib et al. 2012), that the SGC was constructed by fusion of separately-evolved coding fragments (Yarus 2021c), and that code division and re-fusion of evolved fragments was a major factor, especially enabling SGC evolution from one primary source of code assignments (Yarus 2023d). Significantly, 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 (Yarus 2022a). Moreover, fusion produces a prolonged elevation, or "crescendo" of changing codes, all very SGC-like. Within such a crescendo, an optimal code will be found by testing as many competent codes as required to select the SGC (Yarus 2023a, 2022a).

Code fusion implies code convergence. Fig. 4A illustrates early code convergence to majority code assignments. The Figure shows two merging codes: if two sources using differing coding assignments join, fused codes assign triplets ambiguously. This makes a fused code less fit. Such ambiguous fusions become extinct. The smaller number of codes that survive fusion will therefore have more consistent assignments. Fit codes after fusion converge to a homogeneous consensus of the incoming code population. This is the likely route to a coherent SGC (Yarus 2022a). The same mechanism rejects of codes from completely independent origins (Fig. 2, 4A).

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 grouped in an initial set (Fig. 2B, 2C). These effects are easily seen by comparing population properties: 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 stereochemical assignment at full-code frequencies almost excludes alternative assignment, to less than one codon in a subsequent near-complete code population (Fig. 2C).

But full stereochemical intiations differ. Completed sets of potential stereochemical assignments at observed frequencies differ greatly among themselves (Fig. 3, Supplementary Information). 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 substantially different fractions of total SGC functions. The evolutionary path from such initiations to the SGC is therefore of different lengths, and requires varied times (Fig. 3A), allowing different proportions of unlike assignment (Fig. 3B). This conclusion does not require extrapolation: even prior experimental stereochemistry (Stereo assignment) limits alien assignment to less than 2 codons (Fig. 2B) of the 64 possible in full codes. Thus, a persuasive estimate for alternative coding within the stereochemically-determined SGC will likely require a completed stereochemical census (Yarus 2017) for SGC triplets.

The SGC probably does not have a single origin. Present calculations imply that there was an initial era almost entirely stereochemical in character, then a second era potentially permitting other kinds of amino acid/triplet assignment. It thereby explains why a metabolic code origin (Wong 1975) cannot be complete (Amirnovin 1997; Ronneberg et al. 2000) why a completely adaptive origin (Freeland and Hurst 1998) becomes difficult to rationalize (Massey 2019), why a random origin is improbable (Knight et al. 1999), and indeed, why even a stereochemical origin (Yarus 2017) is unlikely to explain every coding assignment (Koonin and Novozhilov 2017). The argument for sequential evolution, linked by the ribonucleopeptide (RNP) transition, follows.

Magnitude of the RNP transition. An 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, the 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 (Longo et al. 2013), rather than the standard complement.

Chemically active, folded forms are a more complex and relevant requirement for the current argument. Chorismate mutase, a 93-amino acid enzyme with a simple alpha-helical structure, has in vivo activity if reduced to only 9 amino acids (Walter et al. 2005). A 213-aa orotidine phosphotransferase (from PRPP) is active if reduced to 13 kinds of amino acids (Akanuma et al. 2002). A reconstructed ancestral bacterial NDP kinase functions with excellent thermal stability and up to 1% control activity with 13 or 14 amino acid alphabets (Shibue et al. 2018). A protein of particular relevance, small ribosomal subunit protein uS8, with 130-aa, binds its RNA composed of 15, 14 or 13 amino acids (Zhao and Akanuma 2023). Notably, though less stable than true thermophilic homologues, reduced uS8 proteins still had high stabilities, half-melted at ~85°C.

In short, 9 to 15 encoded amino acids appear to be required to construct primordial structurally and chemically active peptides hundreds of amino acids long, with a frequent threshold at 13 different amino acids and above. The evident conclusion is that the extrapolated stereochemical encoded complement of 15 amino acids (**Chemistry for a code**, above) is sufficient to encode the RNP transition. Therefore, an initial RNA-based translation apparatus could construct its more capable, more complex successor. This simple succession is depicted in the summary Fig. 4B.

These data therefore suggest that an initial code relying on RNA chemistry encoded and used polypeptides with major structural and enzymatic capabilities. After this ribonucleopeptide (RNP) transition, evolution to modern translation and the full SGC is accessible, requiring extensive refinement (Grosjean and Westhof 2016; Lei and Burton 2022), but not radical evolutionary innovation.

Life's tempo and mode alter after the RNP transition. A characteristic of RNA catalysis is that, though certainly more versatile and faster than one would have imagined, it is still slow, compared to protein catalysis (Yarus 2002). Biocatalysis will accelerate by orders of magnitude after the RNP transition brings peptide groups to bear on reaction transition states. The addition of histidine alone to coding (Majerfeld et al. 2005; Ruta et al. 2007) and peptide (Turk-Macleod et al. 2012) will likely have a notable catalytic impact. As a result, cell activities, including cell division, likely accelerate around the RNP transition. Transitioning cells race away from others with slower, less versatile catalysis, and we are their inheritors.

Fate of early stereochemical assignments after the RNP transition. Crick's intuition about the 'freezing' of the code (Crick 1968) applies particularly to the RNP transition (Fig. 4). Even with dramatic changes attending the appearance of peptidyl aaRS and ribosomes, initial RNA assignments would necessarily be preserved in the RNP era to preserve synthesis of indispensable early peptides (Ardell and Sella 2002). Thus most or all initial stereochemical RNA-era triplet assignments probably survive in the SGC, though the machinery that selects them has changed. Distinguishing RNA-era assignments and RNP or more complex assignments is now clearly worth specific thought.

An intriguing possibility is that amino acids with multiple assigned sets of codons were assigned at two times, in two different ways. Positing Table 1's assignments as RNA-era encodings might be tempting. But this is not plausible; the true path must be more intricate. For example, arginine has two sets of codons - CGN and AGR, has a complex biosynthesis (Xu et al. 2007), probably was encoded later (Trifonov 2000), and exhibits multiple affinities to cognate RNAs (Yarus 1988; Yarus and Christian 1989; Janas et al. 2010; Yarus 2017). It now seems possible that some Arg assignments reflect RNA stereochemistry used for assignment after biosynthesis in the RNP-era, while other triplets reflect RNP or other multi-molecular affinity, possible only in the RNP-era. Accordingly, even complex amino acids with intricate biosynthetic origins can be assigned via affinity for RNA alone. Of course, genuinely persistent RNA-era assignments are also possible (consider Ile, Tab.1), as are assignments unique to post-RNP evolution.

Coding is partial at the RNP transition. As a result of considerations just above, coding will likely be incomplete at the RNP transition, biased toward encoding of first-available and RNA-interacting amino acids. How is the SGC completed?

The RNP transition can make code logic more accepting. Abstractly, coding after the RNP transition might continue to reject foreign assignments, via fusion incompatibility. But though speculative, an appealing hypothesis is that as the biome progressed, endocytotic cellular fusions became more frequent. Thus a code would exist in one, and a different code in another, membrane-bound compartment within the same cell. Long-lived proximity between divergent codes in a post-RNP era can elevate the frequency of successful LGT and thus code transfers.

Similar, but later events are well-studied. During the later DNA era (Fig. 4B), massive multi-directional genetic transfer between compartments was the rule. For example, massive LGT occurred from the alpha-proteobacterial precursor of the mitochondrion to the protoeukarotic nucleus (Leister and Kleine 2011). 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 (Gabaldón and Huynen 2007). Gene transfer between DNA compartments was evidently strongly favored.

DNA appears after the RNP transition. Deoxyribose is not among easily accessible 5-carbon sugars (Furukawa et al. 2019) and the unusual radical catalysis universally required to generate it from ribose (Poole et al. 2014) 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 Darwinian descent itself might have depended on pre-existing peptide encoding/translation (Yarus 2022b) and thus on the RNP transition; expression itself sharply elevated the benefit from accurate inheritance of numerous peptides.

RNA coding history. Fig. 4B attempts to put the code's RNA era in historical context, specifically respecting its intricate history. The RNA world requires the ability to replicate RNAs accurately, though eased by the finding that simplest active translational RNAs are perhaps as small as tetramers, pentamers and hexamers (Chumachenko et al. 2009; Turk et al. 2011; Yarus 2011). The first coding assignments are made using RNA affinities alone (Yarus et al. 2005; Yarus 2017), but as soon as peptides exist, they can join with ribonucleotides to add to the specificity of amino acid binding and triplet assignment (Carter 2015; Tagami et al. 2017). The possibility that other ancient molecules also joined ribonucleotide sequences to form later amino acid binding sites should not be neglected: polar lipids (Khvorova et al. 1999; Vlassov et al. 2001; Janas et al. 2006; Czerniak and Saenz 2022) and condensates (Banani et al. 2017) present other well-characterized RNA associations.

The result is a set of assignments more-than-sufficiently complete to make the RNP transition (**Magnitude of the RNP transition**, above). After the RNP transition, the encoding of substantial peptides enables the appearance of the proto-ribosome, usually modeled as resembling the region of the modern peptidyl transferase (Agmon et al. 2005; Bowman et al. 2020). The assembly of the early code relies on the code division-fusion crescendo which helps the code converge to the SGC, speeds its evolution by allowing parallel progress, and presents a prolonged series of close SGC relatives, facilitating its selection (Yarus 2021c, 2022a, 2023c). Wobble is a late event in code evolution, because accurate wobble requires a large allosteric ribosome (Ogle et al. 2001), a refined tRNA structure (Schultz and Yarus 1994; Dale et al. 2009), and additionally, early wobble greatly hinders code completion (Yarus 2023b). Translation initiation (Schmitt et al. 2019) and termination (Burroughs and Aravind 2019) evolve over a prolonged period, and become late coding events, their evolution continuing after the separation of major biological domains. Near-modern coding capabilities produce the protein aaRS (Ribas de Pouplana 2020), and deoxyribonucleotides (Poole et al. 2014), as well as nucleotide modifying enzymes that refine the codon-anticodon complex (Grosjean and Westhof 2016; Lei and Burton 2022). One complete code becomes universal (Vetsigian et al. 2006) and diverges from LUCA to produce multi-domain modern biota (Crapitto et al. 2022)

Biology as Anthology. This aphorism defines life via its ability to combine advantages evolved separately (Yarus 2023a). Current work resolves a functional ambiguity (Yarus 2023c): efficient code division-and-fusion SGC evolution allows one principal route to the final code to exist, but also seems to offer code fusion from varied origins into one final SGC.

While fusion into the nascent SGC clearly can exist (Fig. 2), it is strictly limited by early stereochemistry, and decreases as chemically-based initial assignment increases (Fig. 1, 2). Thus, fused code evolution both suggests joining independent partial codes, but simultaneously limits successful early joining by the same mechanism (Yarus 2022a).

The historical SGC probably contains a constrained code anthology. It still bears evidence of early stereochemical sources, conserved because assignments must be conserved across the RNP transition to preserve prior peptide capabilities. The potential size of early non-stereochemical RNA-era patches is small, determined by the exact stereochemical history of the code's origin (Fig. 4). The SGC also accepted post-RNP transition assignments, because coding was probably incomplete during the RNP transition. The magnitude of later, non-stereochemical contributions remains to be resolved, but this can be clarified by practical experiments.

Multiple coding eras. Credibly, the SGC significantly evolved in three eras; the RNA world, after the RNP transition, and in the era of the DNA genome, using differing mechanisms. In the RNA era, code fusion restricted and homogenized coding, but in a later era, a different fusion mechanism may have broadened coding. In the DNA era, near-complete codes were compared and a victor selected that became the SGC (Vetsigian et al. 2006), and encoded LUCA's diverse proteins (Mat et al. 2008).

References

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

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

Alkatib S, Scharff LB, Rogalski M, Fleischmann T, Matthes A, Seeger S, Schöttler MA, Ruf S, Bock R. 2012. The Contributions of Wobbling and Superwobbling to the Reading of the Genetic Code. *PLOS Genet* **8**: e1003076.

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

Ardell DH, Sella G. 2002. 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.

Banani SF, Lee HO, Hyman AA, Rosen MK. 2017. Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol* **18**: 285–298.

Bowman JC, Petrov AS, Frenkel-Pinter M, Penev PI, Williams LD. 2020. Root of the Tree: The Significance, Evolution, and Origins of the Ribosome. *Chem Rev* **120**: 4848–4878.

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

Carter CW. 2015. What RNA World? Why a Peptide/RNA Partnership Merits Renewed Experimental Attention. *Life Basel Switz* **5**: 294–320.

Chumachenko NV, Novikov Y, Yarus M. 2009. Rapid and simple ribozymic aminoacylation using three conserved nucleotides. *J Am Chem Soc* **131**: 5257–63.

Ciesiolka J, Illangasekare M, Majerfeld I, Nickles T, Welch M, Yarus M, Zinnen S. 1996. Affinity selection-amplification from randomized ribooligonucleotide pools. *Methods Enzym* **267**: 315–35.

Cleaves HJ. 2010. The origin of the biologically coded amino acids. *J Theor Biol* **263**: 490–498.

Crapitto AJ, Campbell A, Harris A, Goldman AD. 2022. A consensus view of the proteome of the last universal common ancestor. *Ecol Evol* **12**: e8930.

Crick FHC. 1968. The Origin of the Genetic Code. *J Mol Biol* **38**: 367–379.

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

Dale T, Fahlman RP, Olejniczak M, Uhlenbeck OC. 2009. Specificity of the ribosomal A site for aminoacyl-tRNAs. *Nucleic Acids Res* **37**: 1202–1210.

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

Freeland SJ, Hurst LD. 1998. The genetic code is one in a million. *J Mol Evol* **47**: 238–248.

Furukawa Y, Chikaraishi Y, Ohkouchi N, Ogawa NO, Glavin DP, Dworkin JP, Abe C, Nakamura T. 2019. Extraterrestrial ribose and other sugars in primitive meteorites. *Proc Natl Acad Sci* **116**: 24440–24445.

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

Gamow G. 1954. Possible Relation between Deoxyribonucleic Acid and Protein Structures. *Nature* **173**: 318–318.

Gilbert W. 1986. The RNA world. *Nature* **319**: 618.

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

Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, Penswick JR, Zamir A. 1966. Structure of a ribonucleic acid sequence. *Science* **147**: 1462–1465.

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

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

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

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

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

Johnson DB, Wang L. 2010. Imprints of the genetic code in the ribosome. *Proc Natl Acad Sci U A* **107**: 8298–8303.

Kapral TH, Farnhammer F, Zhao W, Lu ZJ, Zagrovic B. 2022. Widespread autogenous mRNA-protein interactions detected by CLIP-seq. *Nucleic Acids Res* **50**: 9984–9999.

Khvorova A, Kwak YG, Tamkun M, Majerfeld I, Yarus M. 1999. RNAs that bind and change the permeability of phospholipid membranes. *Proc Natl Acad Sci U A* **96**: 10649–54.

Knight RD, Freeland SJ, Landweber LF. 1999. Selection, history and chemistry: the three faces of the genetic code. *Trends Biochem Sci* **24**: 241–7.

Koonin EV, Novozhilov AS. 2008. Origin and evolution of the genetic code: The universal enigma. *IUBMB Life* **61**: 99–111.

Koonin EV, Novozhilov AS. 2017. Origin and Evolution of the Universal Genetic Code. *Annu Rev Genet* **51**: 45–62.

Kumar RK, Yarus M. 2001. RNA-catalyzed amino acid activation. *Biochemistry* **40**: 6998-7004.

Lei L, Burton ZF. 2022. "Superwobbling" and tRNA-34 Wobble and tRNA-37 Anticodon Loop Modifications in Evolution and Devolution of the Genetic Code. *Life Basel Switz* **12**: 252.

Leister D, Kleine T. 2011. Chapter three - Role of Intercompartmental DNA Transfer in Producing Genetic Diversity. In *International Review of Cell and Molecular Biology* (ed. K.W. Jeon), Vol. 291 of, pp. 73–114, Academic Press https://www.sciencedirect.com/science/article/pii/B9780123860354000033 (Accessed October 20, 2023).

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

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

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

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

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

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

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

Massey SE. 2019. Genetic Code Error Minimization as a Non-Adaptive But Beneficial Trait. *J Mol Evol* **87**: 4–6.

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

Miller SL. 1953. Production of amino acids under possible primitive earth conditions. *Science* **117**: 528–529.

Ogle JM, Brodersen DE, Clemons WM, Tarry MJ, Carter AP, Ramakrishnan V. 2001. Recognition of Cognate Transfer RNA by the 30S Ribosomal Subunit. *Science* **292**: 897–902.

Ostrov N, Landon M, Guell M, Kuznetsov G, Teramoto J, Cervantes N, Zhou M, Singh K, Napolitano MG, Moosburner M, et al. 2016. Design, synthesis, and testing toward a 57-codon genome. *Science* **353**: 819–822.

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

Poole AM, Horinouchi N, Catchpole RJ, Si D, Hibi M, Tanaka K, Ogawa J. 2014. The Case for an Early Biological Origin of DNA. *J Mol Evol* **79**: 204–212.

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

Rodin AS, Szathmáry E, Rodin SN. 2011. On origin of genetic code and tRNA before translation. *Biol Direct* **6**: 14.

Ronneberg TA, Landweber LF, Freeland SJ. 2000. Testing a biosynthetic theory of the genetic code: fact or artifact? *Proc Natl Acad Sci U A* **97**: 13690–5.

Rubio Gomez MA, Ibba M. 2020. Aminoacyl-tRNA synthetases. *RNA N Y N* **26**: 910–936.

Ruta J, Grosset C, Ravelet C, Fize J, Villet A, Ravel A, Peyrin E. 2007. 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.

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

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

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

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

Shibue R, Sasamoto T, Shimada M, Zhang B, Yamagishi A, Akanuma S. 2018. Comprehensive reduction of amino acid set in a protein suggests the importance of prebiotic amino acids for stable proteins. *Sci Rep* **8**: 1227.

Shulgina Y, Eddy SR. 2021. A computational screen for alternative genetic codes in over 250,000 genomes. *eLife* **10**: e71402.

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

Trifonov EN. 2000. Consensus temporal order of amino acids and evolution of the triplet code. *Gene* **261**: 139–151.

Turk RM, Illangasekare M, Yarus M. 2011. Catalyzed and spontaneous reactions on ribozyme ribose. *J Am Chem Soc* **133**: 6044–50.

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

Vetsigian K, Woese C, Goldenfeld N. 2006. Collective evolution and the genetic code. *Proc Natl Acad Sci U A* **103**: 10696–701.

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

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

Weber AL, Lacey Jr JC. 1978. Genetic Code Correlations: Amino Acids and Their Anticodon Nucleotides. *J Mol Evol* **11**: 199–210.

Weber AL, Miller SL. 1981. Reasons for the Occurrence of the Twenty Coded Protein Amino Acids. *J Mol Evol* **17**: 273–284.

Wilson TJ, Lilley DMJ. 2021. The potential versatility of RNA catalysis. *Wiley Interdiscip Rev RNA* **12**: e1651.

Wong JT-F. 1975. A Co-Evolution Theory of the Genetic Code. *Proc Natl Acad Sci U A* **72**: 1909–1912.

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

Yarus M. 2021a. Crick Wobble and Superwobble in Standard Genetic Code Evolution. *J Mol Evol* **89**: 50–61.

Yarus M. 2021b. Evolution of the Standard Genetic Code. *J Mol Evol* **89**: 19–44.

Yarus M. 2021c. Fitting the standard genetic code into its triplet table. *Proc Natl Acad Sci U S A* **118**: e2021103118.

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

Yarus M. 2022b. Evolution and change: a principle of least selection. 2021.06.27.450095. https://www.biorxiv.org/content/10.1101/2021.06.27.450095v3 (Accessed December 23, 2022).

Yarus M. 2023a. Order of events in a developing genetic code. 2022.12.31.522385. https://www.biorxiv.org/content/10.1101/2022.12.31.522385v1 (Accessed February 3, 2023).

Yarus M. 2023b. Order of events in a developing genetic code. 2022.12.31.522385. https://www.biorxiv.org/content/10.1101/2022.12.31.522385v1 (Accessed April 26, 2023).

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

Yarus M. 2023d. The genetic code assembles via division and fusion, basic cellular events. 2023.05.01.538992. https://www.biorxiv.org/content/10.1101/2023.05.01.538992v2 (Accessed June 14, 2023).

Yarus M. 1988. A specific amino acid binding site composed of RNA. *Science* **240**: 1751–1758.

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

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

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

Yarus M. 2017. The Genetic Code and RNA-Amino Acid Affinities. *Life* **7**: 13.

Yarus M, Christian EL. 1989. Genetic Code Origins. *Nature* **342**: 349–350.

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

Yuan J, O'Donoghue P, Ambrogelly A, Gundllapalli S, Sherrer RL, Palioura S, Simonović M, Söll D. 2010. Distinct genetic code expansion strategies for selenocysteine and pyrrolysine are reflected in different aminoacyl-tRNA formation systems. *FEBS Lett* **584**: 342–349.

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

Legends

Table 1 – Cognate codons and anticodons in RNA-amino acid binding sites. Description, references and details are found in (Yarus 2017). 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 to later SGC 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 incompatible 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 triplets at a time. The number of accepted ALT-like assignments is plotted versus the number of triplets potentially transferred. 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_code 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 (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 **Chemistry for a code** in the text 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 – **A 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 (Yarus 2021b, 2023d) 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 one 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 a randomly chosen code 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 for late coding here), Pdivn is the probability/passage that a random existing coding table will divide identically (Yarus 2021b, 2023d).

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 (see 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, SGC-like 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 (see supplementary information) which has no assignments in common with the SGC. In addition, SGC-like 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 (see supplementary information) which has no assignments in common with the SGC. In addition, SGC-like 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 information files. 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 (**Chemistry for a code**, in text; 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 identical 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

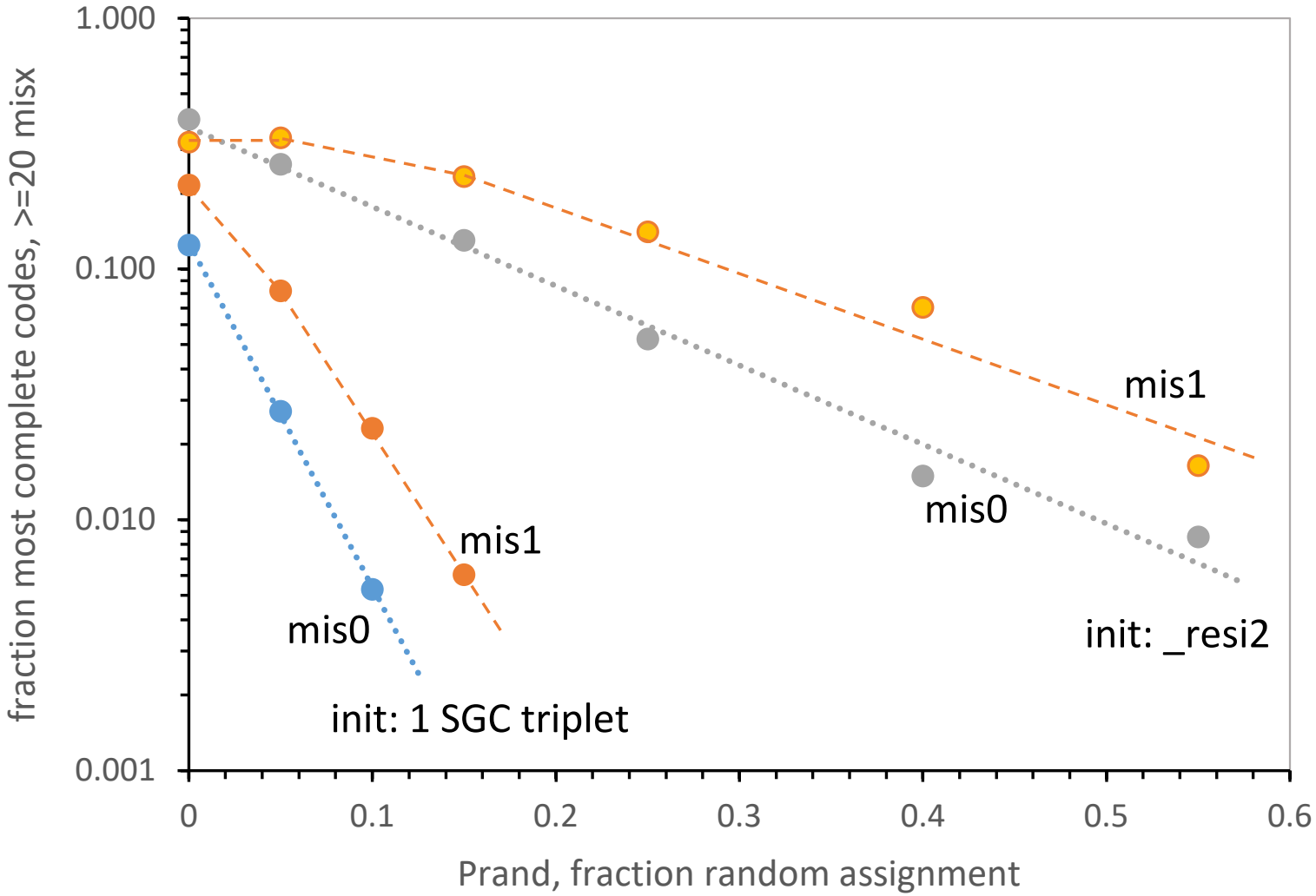
1.000
0.100
0.010
0.001
fraction most complete codes, >=20 misx
0
0.1
0.2
0.3
0.4
0.5
0.6
Prand, fraction random assignment
mis1
mis0
mis1
mis0
init: _resi2
init: 1 SGC triplet

Figure 1

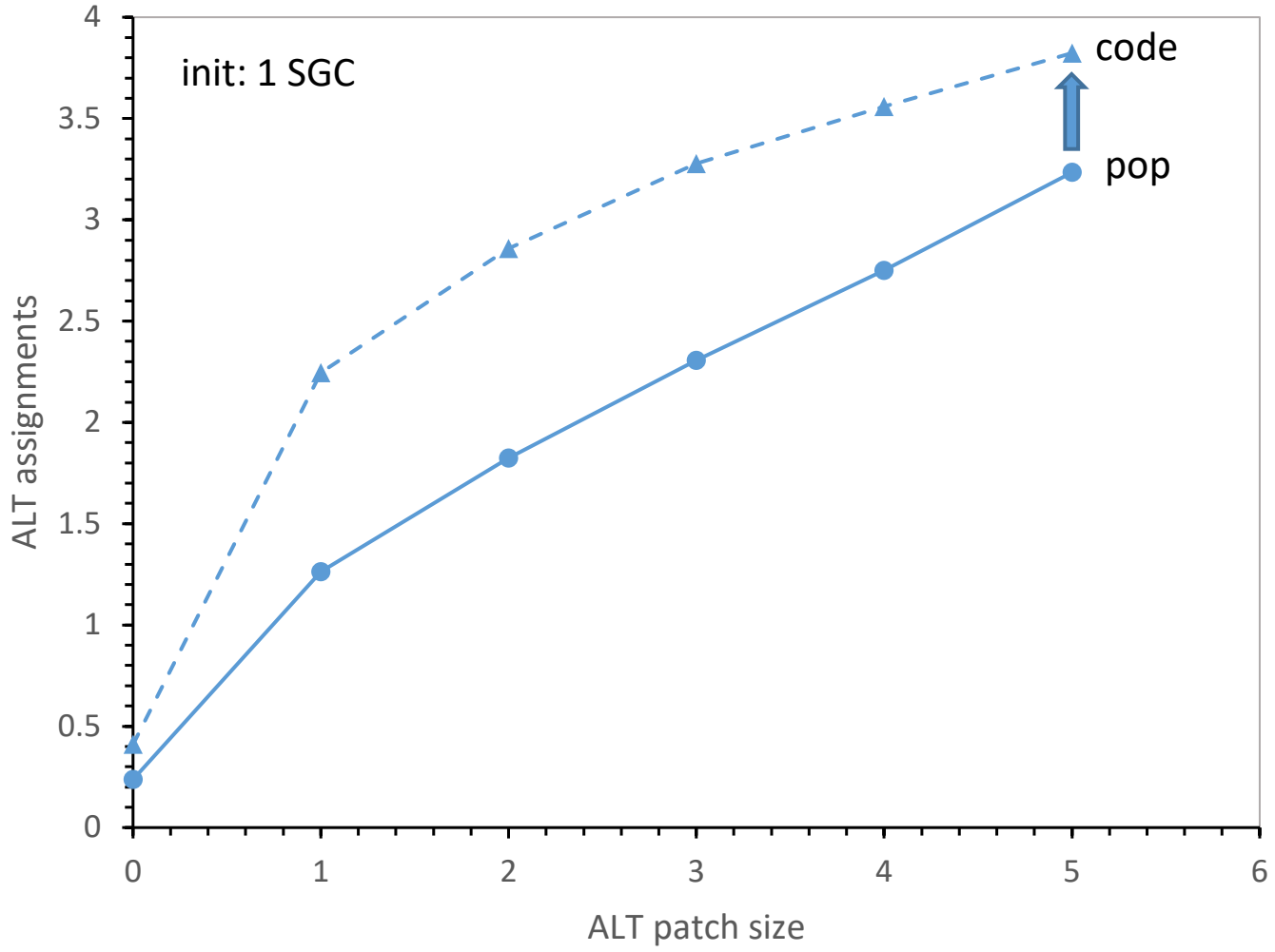
init: 1 SGC
code
pop
ALT assignments
ALT patch size
0
0.5
1
1.5
2
2.5
3
3.5
4
0
1
2
3
4
5
6

Figure 2A

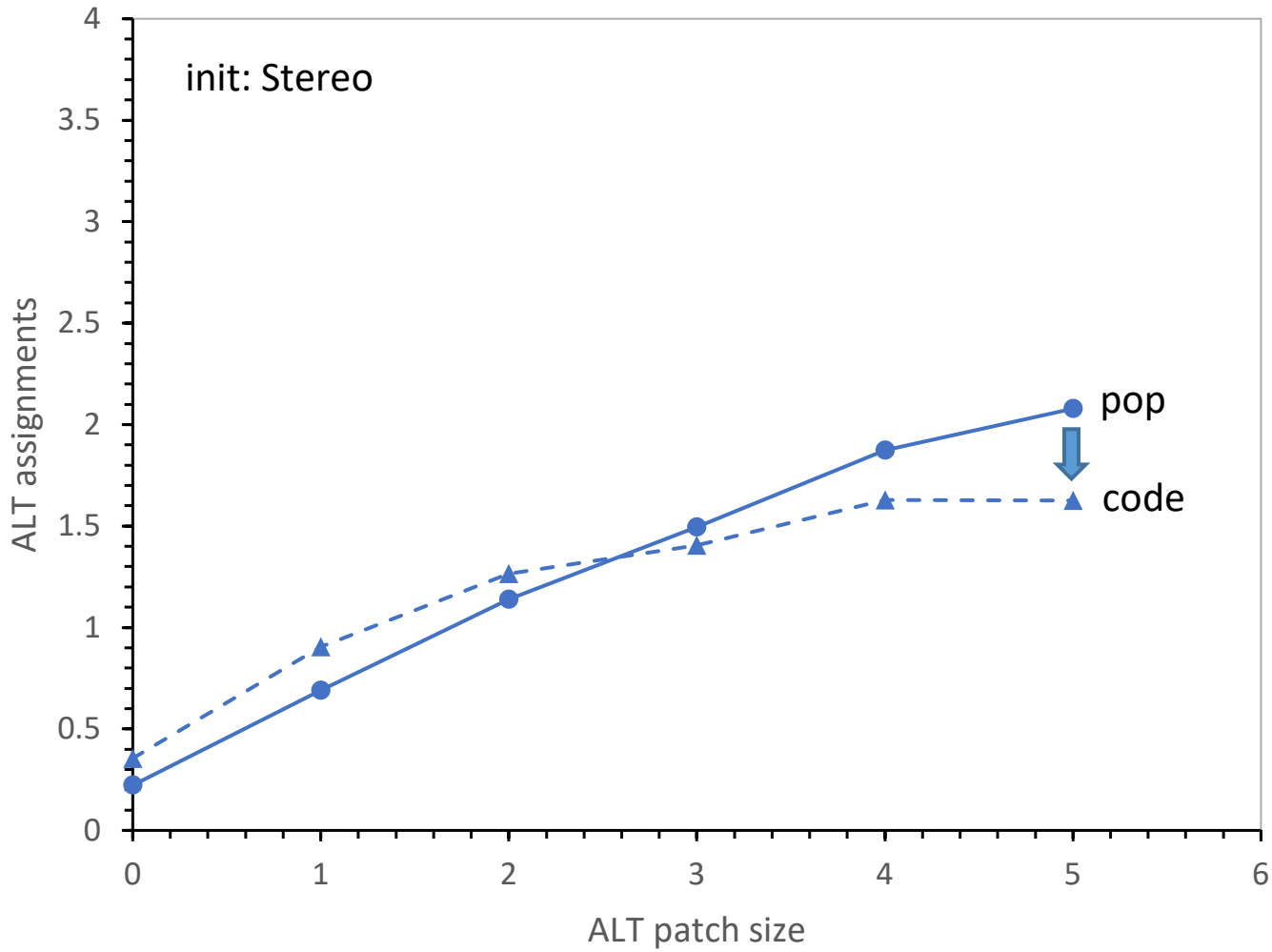
init: Stereo
pop
code
ALT assignments
ALT patch size
0
0.5
1
1.5
2
2.5
3
3.5
4
0
1
2
3
4
5
6

Figure 2B

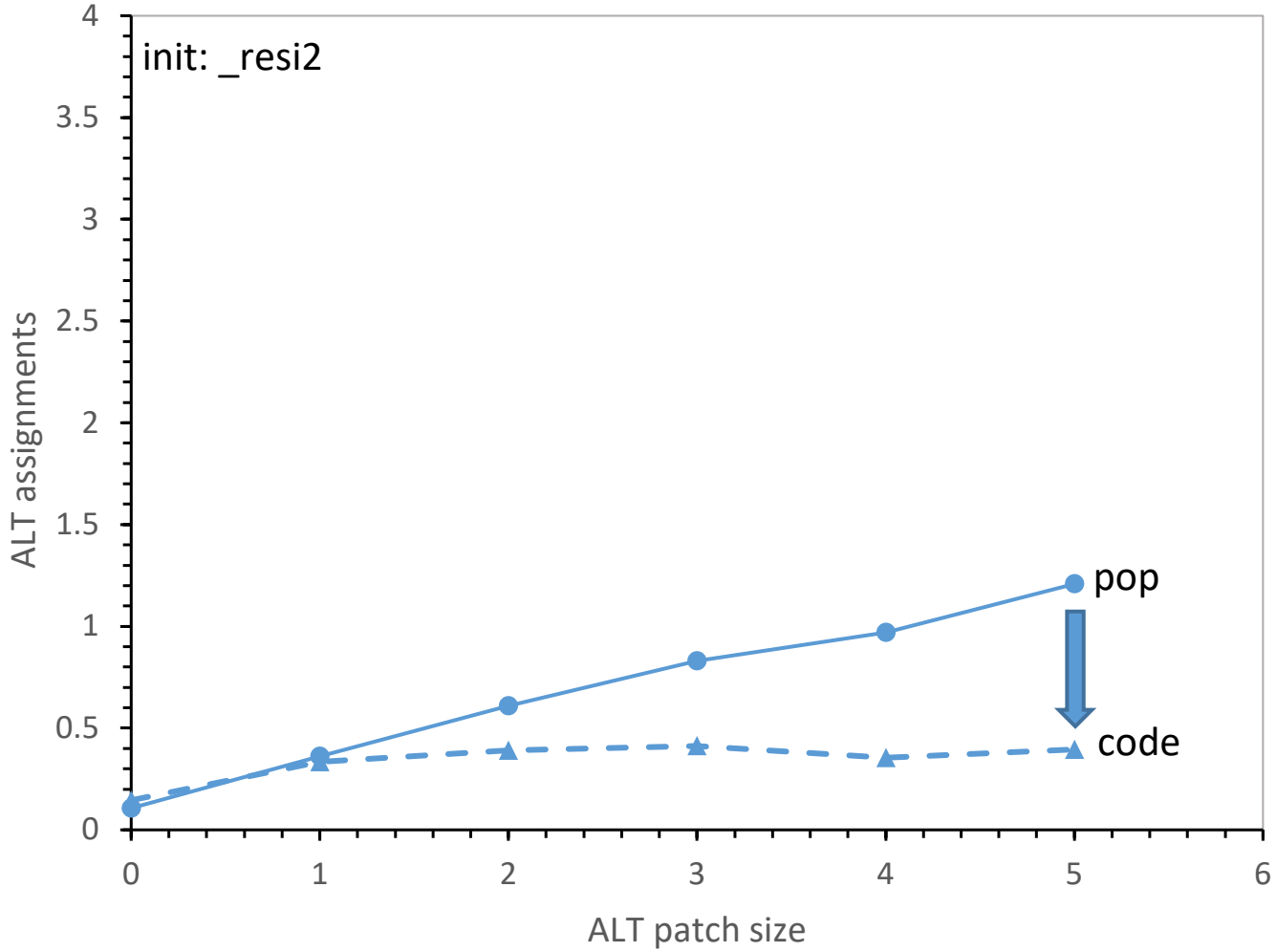
init: _resi2
ALT assignments
ALT patch size
pop
code
0
0.5
1
1.5
2
2.5
3
3.5
4
0
1
2
3
4
5
6

Figure 2C

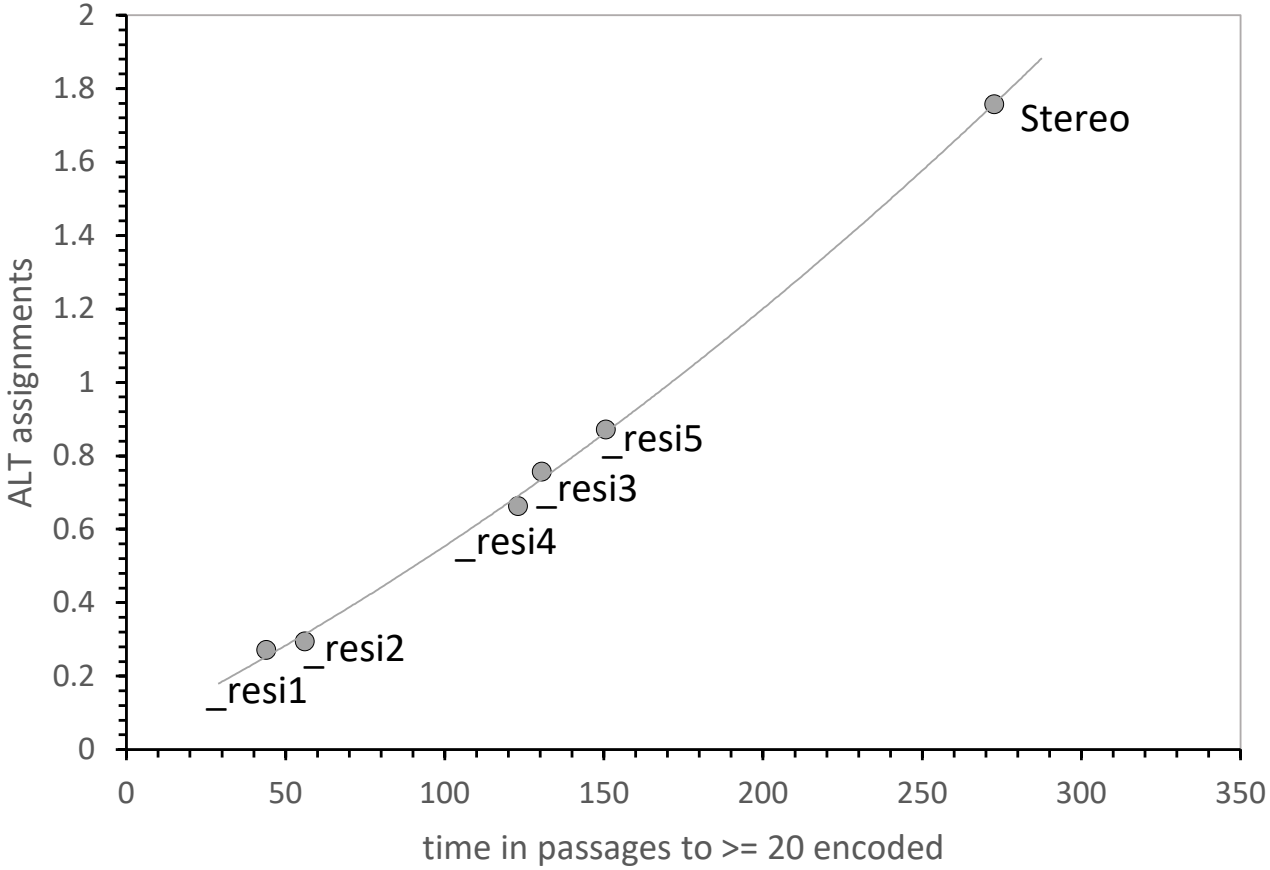
ALT assignments
time in passages to >= 20 encoded
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
0
50
100
150
200
250
300
350
Stereo
_resi5
_resi3
_resi4
_resi2
_resi1

Figure 3A

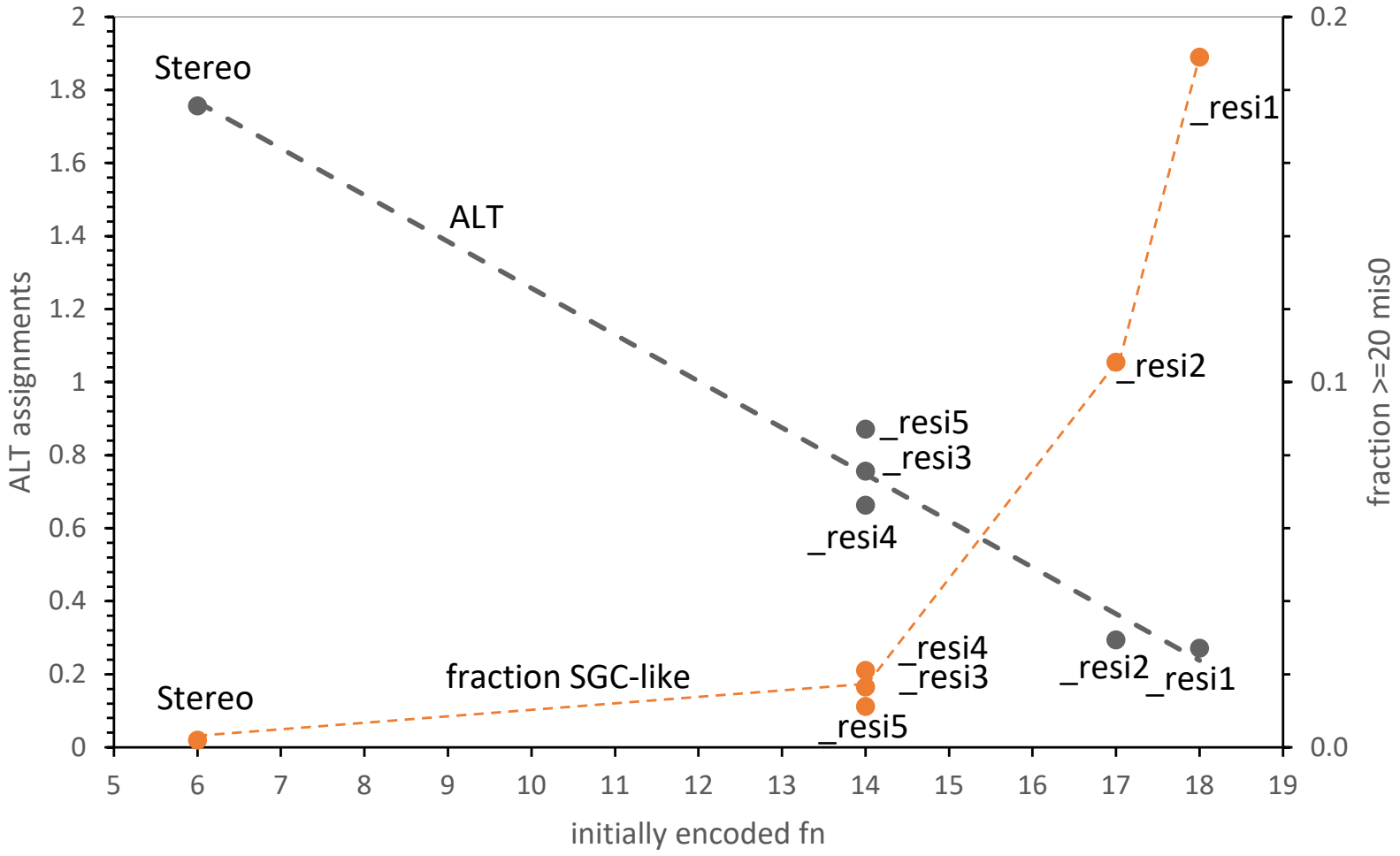
ALT assignments
fraction >=20 mis0
initially encoded fn
Stereo
ALT
_resi1
_resi2
_resi5
_resi3
_resi4
fraction SGC-like
Stereo
_resi4
_resi3
_resi5
_resi2
_resi1
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0.2
0.1
0.0
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

Figure 3B

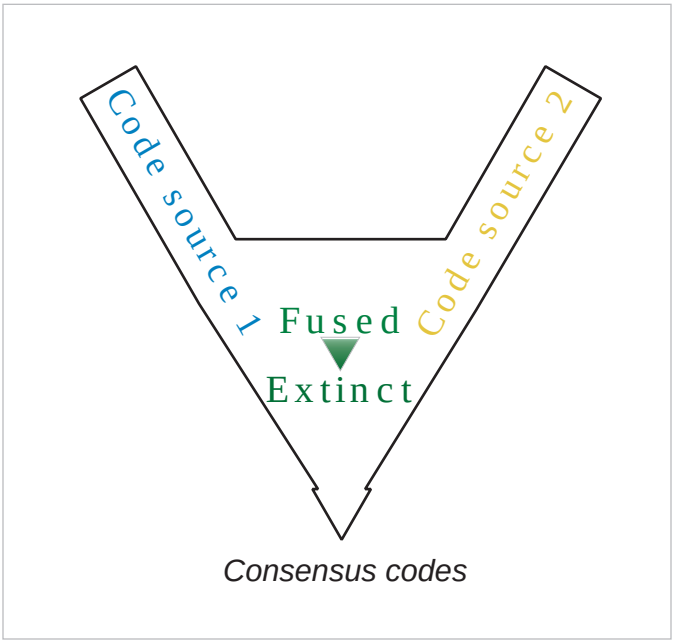

Figure 4A

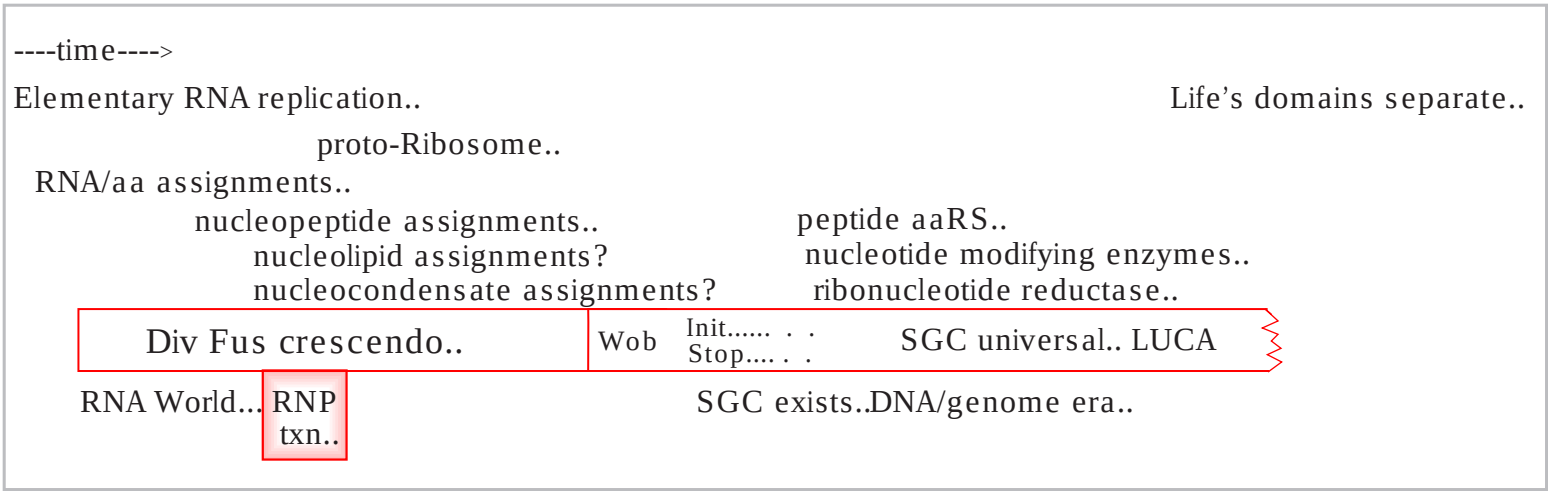
----time---->
Elementary RNA replication..
Life's domains separate..
proto-Ribosome..
RNA/aa assignments..
nucleopeptide assignments..
peptide aaRS..
nucleolipid assignments?
nucleotide modifying enzymes..
nucleocondensate assignments?
ribonucleotide reductase..
Div Fus crescendo..
Wob
Init.....
Stop....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

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