

The Early Evolution of the Genetic Code

Minireview

Robin D. Knight* and Laura F. Landweber*
Department of Ecology and Evolutionary Biology
Princeton University
Princeton, New Jersey 08544

Evolutionary inferences rely on diversity. The source of differences among organisms is accumulated divergence from a common ancestor, which may be random or selected. When a system is adaptive yet highly complex, one can follow its evolution from a simpler state in one of two ways: from fossilized transitional forms, or from early-diverging extant organisms. This is how, for example, we can trace the evolution of trichromatic vision in primates or flowers in angiosperms.

The problem becomes harder when no intermediate states exist. In particular, hypotheses about evolution prior to the Last Universal Common Ancestor of extant life (LUCA) defy standard techniques. Biochemical pathways do not fossilize, precluding direct inferences about ancestral states, and by definition no lineages diverging before the LUCA survive. Thus the diversity of extant life reveals little about general principles, since biochemical necessities mingle with quirks inherited from the shared ancestor. Consequently, it is difficult to explain why highly conserved and universal systems such as the translation apparatus are the way they are.

Early Evolution of the Code: Extraordinary Techniques for Extraordinary Problems

In the absence of evidence, many of the most interesting questions about the genetic code have fallen into a twilight zone of speculation and controversy. Although it is generally accepted that the modern code evolved from a simpler form, there has been no consensus about when the initial code evolved or what it was like, how and when particular amino acids were added, how and when the modern tRNA/synthetase system arose, or the processes by which the code could have expanded. Now, detailed study of the components of the translation apparatus is at last making these questions tractable.

Three general approaches have recently yielded surprising intimations about how the genetic code evolved. The first is to appeal to general principles at a primary level, in this case the chemistry of nucleic acids and amino acids, to infer how a translation system might be constrained. The second is to alter parts of the translation apparatus *in vitro* in ways that might reflect earlier states, showing what changes are possible. The third is to examine the phylogeny of particular components, revealing how they have changed since the LUCA (or, in the case of paralogous genes, even before the LUCA), and to extrapolate backward from the principles thus revealed. Here we show how key applications of these approaches begin to provide a general framework for understanding the origin and development of the code.

Amino Acid/Nucleotide Interactions in the RNA World and Earlier

Because RNA is unstable and difficult to synthesize, the first genetic material may have used a simpler backbone than ribose. One candidate is peptide nucleic acid (PNA), in which the backbone is polymeric N-(2-aminoethyl)glycine (AEG) and the N-acetic acids of the bases (N₂ for purines, N₁ for pyrimidines) are linked via amide bonds (Figure 1). This is an attractive scenario because AEG forms in spark-tube experiments that also produce amino acids (Nelson et al., 2000), and may spontaneously polymerize at 100°. The N-acetic acids of the bases are also accessible in prebiotic syntheses, which suggests that PNA could have been an early genetic material (although the evidence is far from conclusive).

The prebiotic plausibility of PNA implies that amino acids and a genetic system based on purines and pyrimidines could have been coproduced and then coevolved on the early earth. Does the genetic code in modern organisms reflect such ancient interactions, or have all traces been erased by subsequent evolution of the translation apparatus?

One can approach this question statistically, asking whether chemistry has influenced codon assignments. SELEX, the selective amplification of nucleic acid molecules that perform particular tasks, can identify specific RNA sequences that bind amino acids (Connell et al.,

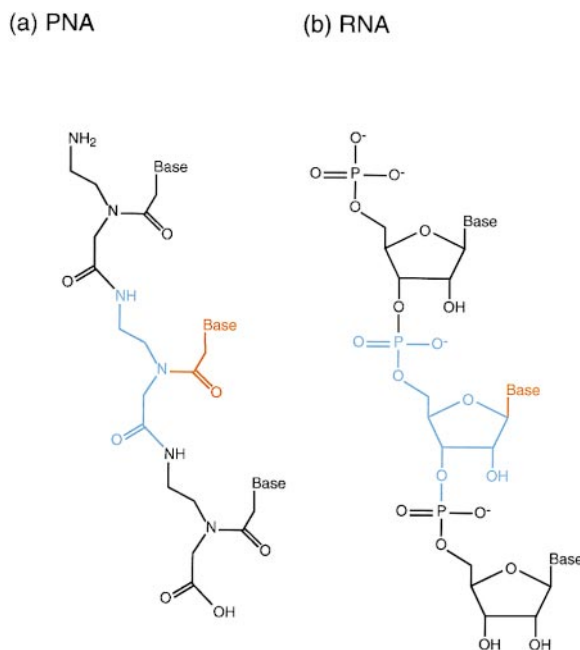


Figure 1. PNA or RNA First?

PNA (a) has a peptide backbone instead of the sugar-phosphate backbone of RNA (b). Unlike ribose, N-(2-aminoethyl)-glycine is formed at high yields under prebiotic conditions and spontaneously produces a stable polymer. However, its uncharged rigid backbone may limit possibilities for catalysis. One backbone monomer is highlighted in blue, with the informational unit highlighted in red.

*E-mail: rdknight@princeton.edu (R. D. K.), lfl@princeton.edu (L. F. L.)

1993). Thus, one can test whether codons that specify a particular amino acid in the canonical genetic code occur disproportionately often at RNA sites that bind it. For instance, we found that arginine binding sites are predominantly composed of Arg codons, even in aptamers selected in different labs using different protocols (Knight and Landweber, 1998).

The recent isolation of aptamers to tyrosine, which bear more Tyr codons than expected at their binding sites (Mannironi et al., 2000), prompted a debate in the April issue of *RNA* over the robustness and interpretation of the statistical evidence for such associations. Yarus extends this analysis to other amino acids for which aptamers are now available (arginine, isoleucine, and tyrosine), and concludes that the overall probability that the observed codon/binding site association would occur by chance is 3.3×10^{-7} (Yarus, 2000). Ellington et al. raise methodological concerns, showing that the choice of statistical techniques and sequences can affect the level of significance of the association (Ellington et al., 2000). We test the robustness of the result by examining all possible combinations of sequences for binding site associations with each codon set (Knight and Landweber, 2000), and show that Arg codons alone significantly associate with arginine binding sites.

Taken together, these papers show that the amino acid aptamers that have been structurally characterized do overrepresent their cognate codons at their binding sites. Although as Ellington et al. point out there are grounds for caution (the structurally characterized aptamers could be a nonrepresentative sample, and the relationship does not hold independently for each of the four nucleotides), we can tentatively conclude that amino acid binding sites are preferentially enriched in certain trinucleotides, which correspond curiously to modern codon assignments. Even DNA aptamers for arginine showed significant codon/binding site associations (Knight and Landweber, 2000), indicating that the backbone is not critical; this is consistent with the idea that an alternative backbone, such as PNA, might have been the original genetic molecule, and may suggest that elements of the modern genetic code predate the RNA world.

If the genetic code evolved from a simpler form, stereochemistry may have produced an initial code that later expanded. The modern code appears highly optimized for resistance to various types of error (Freeland et al., 2000), which complicates the situation. How could stereochemical codon assignments reflect the same properties that affect substitution rates of amino acids within proteins? Clearly, more aptamer data are needed to establish the generality of codon/binding site associations, and to assess the relative roles of chemistry and selection in shaping the earliest genetic codes.

The Origins of Aminoacyl-tRNA Synthesis

Although protein enzymes catalyze tRNA aminoacylation today, they cannot have existed before protein synthesis itself. It is widely accepted that ribozymes predated proteins, and several labs have recently isolated ribozymes with peptidyltransferase activity. This shows that specific peptide synthesis could have arisen in an RNA world. Two ribozymes of particular interest come from the Yarus lab and the Szostak lab.

Illangasekare and Yarus selected a self-aminoacylating ribozyme using Phe-AMP as a substrate (Illangasekare and Yarus, 1999a). One 95-mer from this pool was highly specific for Phe, accelerating the reaction 6×10^7 -fold over background and preferring Phe-AMP 10^4 -fold over other aminoacyladenylates. This compares favorably to yeast PheRS on both counts, indicating that RNA can catalyze aminoacylation at least as well as do proteins. A 29 nt aminoacylating RNA was later constructed (Illangasekare and Yarus, 1999b). Although this tiny ribozyme is not specific for any amino acid, it can catalyze peptide bond formation as well, suggesting that both these reactions may have been easily accessible to RNA world metabolisms.

Lee et al. selected a self-aminoacylating ribozyme using two different substrates: first, a hexanucleotide complementary to a 3' guide sequence and derivatized with Phe-biotin, and second, cyanomethyl-activated glutamine (Lee et al., 2000). This produced 170 nt "ambidextrous" ribozymes with two independent active domains. Because the transfer from a 5'-OH to a 3'-OH is energetically neutral, a ribozyme that catalyzes transfer from the 3'-OH of another molecule to its own 5'-OH should also perform the reverse reaction if its 5'-OH is already aminoacylated. Thus, when provided with Gln-CME, the ambidextrous ribozymes aminoacylated tRNA molecules that bound the guide sequence. Although these ribozymes are not as fast or selective as that isolated in the Yarus lab, they can specifically aminoacylate tRNA *in trans*, as do modern synthetases.

Because several aminoacylation specificities (Lys, Gly, possibly Tyr/Trp) appear to have evolved several times in independent lineages, it may also be relatively easy for proteins to evolve aminoacyl-tRNA synthetase (aaRS) activity. Chihade and Schimmel attempted to reconstruct a primitive aaRS by linking a minimal aminoacyladenylate-forming domain of Ala-RS to a nonspecific RNA binding domain (Chihade and Schimmel, 1999). Although the resulting protein could aminoacylate a tRNA-Ala-derived microhelix at rates comparable to an aaRS that permits cell growth in yeast, this construct was still large: over 600 amino acids long. Thus, protein synthesis was presumably highly developed by the time protein aaRSs began to replace ribozymes.

Aminoacyl-tRNA Synthetases and the Expanding Code

Were all 20 amino acids in our genetic code present in the RNA world? If so, ribozymes must catalyze a tremendous range of reactions; alternatively, the RNA world might have relied on the few prebiotically available amino acids. Although the aaRSs within each class are related to each other, and hence arose by duplication and divergence of two original synthetases, these duplications could reflect either addition of new amino acids to the code or takeover of existing amino acids from ribozyme synthetases. Although SELEX may suggest an ancient stereochemically determined code, the intermediate transitions are unclear.

New amino acids may have been initially synthesized from metabolic precursors by tRNA-dependent processes, with synthetases capable of directly charging them to tRNAs evolving only later. The new synthetase would capture some of the tRNAs, and hence some of the codons, of its ancestor, assigning metabolically



Figure 2. Atavistic GlnRS

GlnRS (Q) → GluRS (E) from *E. coli* and *H. sapiens*, shown with an example of GluRS from each of the three domains. Residues conserved across both specificities are highlighted in green; those conserved only across GlnRS are highlighted in blue, and those conserved but differing between GlnRS and GluRS are highlighted in yellow. The changes that remove Gln specificity are marked in red. Note that different residues changed in the two cases: this may be because the *E. coli* experiment selected against efficient mischargers, since mischarging of wild-type tRNA^{Gln} with Glu inhibited protein synthesis.

related amino acids to adjacent codons (Wong, 1981). This type of code expansion requires that aaRSs acquire new specificities. Although suppressor mutants are typically altered tRNAs, and never aaRSs, two recent studies show that aaRSs can be engineered to retrace their evolutionary history.

Although all organisms have a dedicated GluRS, GlnRS appears to have arisen as a paralog of GluRS in eukaryotes, with subsequent lateral transfer to a few other lineages. Most bacteria and archaea use GluRS to charge tRNA-Gln with glutamate, and then convert it to glutamine on the tRNA by a transamidase. Agou et al. analyzed a structure-based alignment of GluRS and GlnRS from different taxa, and identified two residues invariant in all GlnRS but absent from GluRS (Agou et al., 1998). Altering these residues to match eukaryotic GluRS reduced selectivity for Gln more than 10,000-fold.

This rational mutagenesis approach is limited to testing the effects of a few mutations. Hong et al. instead randomized sections of GlnRS and selected the variants best conferring GluRS specificity *in vivo*, using *E. coli* GlnRS as a starting point (Hong et al., 1998). Two changes, though interestingly different from the ones noted in Agou et al., improved Glu recognition 3- to 5-fold (Figure 2). This GluRS was inefficient, probably because it mischarged wild-type tRNA^{Gln} with Glu. Combining both approaches by mutating and selecting an orthogonal tRNA/synthetase pair that does not affect the components already in the cell, such as human GlnRS and tRNA-Gln in *E. coli*, might allow a more complete identity switch.

To add amino acids to the code, the original aaRS must relinquish some of its isoacceptor tRNAs to its new paralog. Li et al. take the first steps toward achieving this process experimentally in an insertion mutant of *E. coli*, LeuRS, which prefers one tRNA-Leu isoacceptor 3-fold

over another instead of charging both at equal rates (Li et al., 1999). In nature, a similar process has taken place in *Thermus thermophilus*, which has two independent pathways for tRNA asparaginylation (Becker et al., 2000). The first is direct formation of Asn-tRNA^{Asn} by an archaeal-type AsnRS; the second is indirect formation, first producing Asp-tRNA^{Asn} by a eubacterial-type AspRS and then transamidating this aminoacyl-tRNA to Asn-tRNA^{Asn}. *T. thermophilus* also has an archaeal-type AspRS, which recognizes and aspartylates only tRNA-Asp, in contrast to the eubacterial AspRS, which recognizes and aspartylates tRNA-Asn as well. Clearly, AspRS has lost the ability to recognize a subset of its tRNA substrates in lineages that have an independent AsnRS. This may indicate that Asn was a relatively recent addition to the code, perhaps postdating the origin of most aaRSs.

Recent Code Evolution: Release Factors and Modified Bases

Thus far we have covered processes that led to the code in the LUCA but did not contribute to its subsequent diversification. Recent variant codes are predominantly changes in a few tRNAs and release factors. Examples of the former are numerous, and are often changes in RNA editing or base modification at the anticodon rather than mutations in the anticodons of tRNA genes themselves. For example, Met is encoded by AUG alone in the standard code, but by AUA and AUG in metazoan mitochondria. tRNA^{Met} normally has anticodon CAU: a mutation to UAU would allow recognition of both A and G at the third codon position by wobble pairing. This would seem the easiest way to effect this change, as UNN anticodons commonly read NNR 2-codon sets. However, *Drosophila*, bovine, and squid tRNA^{Met} instead retain the CAU anticodon sequence but modify the C to 5-formylcytidine, which recognizes both A and G

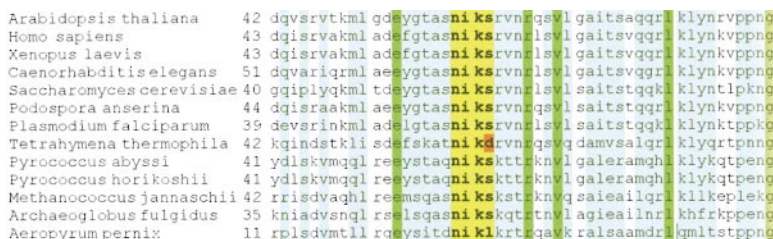


Figure 3. Comparison of eRF1 Homologs from Different Taxa

Release factors are highly conserved compared to aaRSs (see Figure 2 for comparison). Highly conserved residues (>50% identity) are blue, and absolutely conserved residues are green. The NIKS motif, yellow, is involved in stop codon recognition and is conserved except in *Tetrahymena* (nonconservative Ser → Asp, red). Interestingly, of the species shown only *Tetrahymena* uses a noncanonical set of termination codons.

(Tomita et al., 1999). Changes in base modification may indeed be a widespread mechanism of producing alternative genetic codes.

Stop codons are the most labile, changing independently in many lineages. This mutability may reflect their rarity—occurring only once per reading frame—or the ease of losing or altering release factors. The sequence of the release factor eRF1 from *Tetrahymena* (Karamyshev et al., 1999), which uses UAA and UAG for Gln instead of stop, may illuminate this question, since the crystal structure of human eRF1 was recently solved (Song et al., 2000) and several homologs are available from other eukaryotes and archaea. Thus, we can form a specific hypothesis about the molecular basis for this change. The NIKS domain is universally conserved and involved in codon recognition, and mutations immediately adjacent to it produce a universal suppressor. However, in *Tetrahymena*, the Ser undergoes a nonconservative mutation to Asp, which may generate the new specificity (Figure 3). Examination of other ciliate, diplomonad, and algal lineages, with parallel changes in termination, will indicate whether this residue is universally important in stop codon recognition.

Conclusions

Together, research into different components of the translation apparatus is beginning to paint a consistent picture of how the genetic code might have evolved. The primordial code, influenced by direct interactions between bases and amino acids probably dates back to the RNA world or earlier. The invention of tRNAs and ribozyme-based aaRSs made this mapping indirect, allowing swapping of amino acids between codons and hence a level of optimization. Additionally, the code probably underwent a process of expansion from relatively few amino acids to the modern complement of 20. By the time protein aaRSs took over, translation was probably well developed; however, some amino acids, such as Gln, Asn, and Trp, may postdate the first protein aaRSs. Today, laboratory experiments that alter the specificity of aaRSs for amino acids and/or tRNA isoacceptors recapitulate some of these processes. Finally, changes to both tRNAs and release factors produced the range of modern codes, particularly through post-transcriptional base modification and changes in release factors. This diversity of events suggests that an explanation for the fixation of the canonical code in the LUCA will require more historical reconstruction than reasoning from chemical principles.

Selected Reading

- Agou, F., Quevillon, S., Kerjan, P., and Mirande, M. (1998). *Biochemistry* 37, 11309–11314.
- Becker, H.D., Roy, H., Moulinier, L., Mazauric, M.H., Keith, G., and Kern, D. (2000). *Biochemistry* 39, 3216–3230.
- Chihade, J.W., and Schimmel, P. (1999). *Proc. Natl. Acad. Sci. USA* 96, 12316–12321.
- Connell, G.J., Illangsekare, M., and Yarus, M. (1993). *Biochemistry* 32, 5497–5502.
- Ellington, A.D., Khrapov, M., and Shaw, C.A. (2000). *RNA* 6, 485–498.
- Freeland, S.J., Knight, R.D., Landweber, L.F., and Hurst, L.D. (2000). *Mol. Biol. Evol.* 17, 511–518.
- Hong, K.W., Ibba, M., and Soll, D. (1998). *FEBS Lett.* 434, 149–154.
- Illangsekare, M., and Yarus, M. (1999a). *Proc. Natl. Acad. Sci. USA* 96, 5470–5475.

- Illangsekare, M., and Yarus, M. (1999b). *RNA* 5, 1482–1489.
- Karamyshev, A.L., Ito, K., and Nakamura, Y. (1999). *FEBS Lett.* 457, 483–488.
- Knight, R.D., and Landweber, L.F. (1998). *Chem. Biol.* 5, R215–R220.
- Knight, R.D., and Landweber, L.F. (2000). *RNA* 6, 499–510.
- Lee, N., Bessho, Y., Wei, K., Szostak, J.W., and Suga, H. (2000). *Nat. Struct. Biol.* 7, 28–33.
- Li, T., Li, Y., Guo, N., Wang, E., and Wang, Y. (1999). *Biochemistry* 38, 9084–9088.
- Mannironi, C., Scerch, C., Fruscoloni, P., and Tocchini-Valentini, G.P. (2000). *RNA* 6, 520–527.
- Nelson, K.E., Levy, M., and Miller, S.L. (2000). *Proc. Natl. Acad. Sci. USA* 97, 3868–3871.
- Song, H., Mugnier, P., Das, A.K., Webb, H.M., Evans, D.R., Tuite, M.F., Hemmings, B.A., and Barford, D. (2000). *Cell* 100, 311–321.
- Tomita, K., Ueda, T., Ishiwa, S., Crain, P.F., McCloskey, J.A., and Watanabe, K. (1999). *Nucleic Acids Res.* 27, 4291–4297.
- Wong, J.T.-F. (1981). *Trends Biochem. Sci.* 6, 33–36.
- Yarus, M. (2000). *RNA* 6, 475–484.