A genetic code from RNA chemistry: binding sites for amino acids and peptides

Michael Yarus (Michael.Yarus@colorado.edu); Department of Molecular, Cellular, and Developmental Biology, University of Colorado Boulder; Boulder, CO, USA

The genetic code is a set of associations, linking the twenty-some common amino acids and the 64 triplet sequences of four nucleotides. These connections are almost universally conserved among life on Earth. The same code, with only small and temporary variations, has apparently been used by every terran organism since before their last universal common ancestor (LUCA). Accordingly, there is a large literature musing, speculating, or guessing about the code's unanimous acceptance by the only biota known on our planet. This short review, however, takes a different tack, suggesting that there is an experimental, currently demonstrable, chemical connection between (some) amino acids and (some) encoding nucleotides.

A molecular fossil? The story begins 23 years ago, with the elucidation of the active site of the *Tetrahymena* group I self-splicing RNA. The group I intron binds a guanosine or G nucleotide cofactor at a specific site to initiate its RNA splicing sequence. Within the same binding site, arginine can also be bound (1). Strikingly, there are conserved codons (CGA and AGA) for arginine at the site of the bound amino acid. The potential implications are clear: perhaps the elusive connection between coding triplets and amino acids was a chemical one, and the nucleotide code was composed by assembling escaped fragments of ancient RNA-amino acid binding sites.

A generalizable test. The means for a general test of this idea came along when selection-amplification (systematic evolution of ligands by exponential enrichment, SELEX) was devised. SELEX combines molecular biology techniques so that new RNA activities, initially present in a tiny minority, can be purified from starting populations of 10^{13} to 10^{15} different randomized ribonucleotide sequences. Among the early amino acid binding sites recovered in this way were new sites for arginine, which repeated the natural example of the self-splicing RNA. They contained unexpected excesses of conserved coding triplets for the amino acid.

The most recent review of RNA amino acid binding sites (2) contains details of 337 independently selected sites containing 18,551 nucleotides in sequences closely associated with 8 chemically varied (aromatic, aliphatic, polar, and charged), bound amino acids. This large sample of data from newly selected binding sites decisively reinforces the original observation: RNA binding sites tend to include cognate coding triplets (mostly anticodons, but some codons). This is much truer of binding sites, for example, than for control sequences, which are nearby, and only accompany binding sites through the SELEX procedure. Clear though the tendency is, some important questions remain. After some new evidence below, I will return to the meaning of this set of selected sequences.

An RNA template. As one possible reaction to these data, I have suggested that RNA sequences originally ordered amino acids on their own surfaces, encoding an amino acid sequence in the sequence of binding nucleotides, which we know (above) can contain coding triplets. The RNA with the sequence of amino acid sites on its surface was called a **D**irect **R**NA **T**emplate, or DRT. This particular use of RNA binding sites was selected because it seemingly represents the simplest coding system. With only two molecular components (activated amino acids and the DRT), one can synthesize peptides with a desired sequence. To clarify matters, it is necessary to keep in mind that the DRT is a hypothetical way of embodying experiments now so numerous and varied that their implications are unequivocal. Small RNAs do fold to form quite specific binding sites for varied amino acids; I, for one, will be very surprised if this is entirely irrelevant to the history of the genetic code. However, we may choose to use this demonstrated chemical capability to build a DRT, which is a hypothesis among other hypotheses about the role of amino acid binding sites made of RNA.

One objection to a DRT focuses on the fact that amino acids are small, and nucleotides are ca. 3 times as big. Therefore, the 18-20 nucleotides which are packed in space specifically to make a pocket for an amino acid are very big with respect to the object bound. Thus, it was dubious, some argued, that two bowling balls (sites) could be brought together so as to oppose two ping-pong balls (amino acids) inside them, which had to be joined by a yet smaller peptide bond (N-C, 1.33 Angstroms).

A selected RNA site for peptide (His-Phe). Such questions can be answered by construction of an RNA site for a peptide (3). We chose to select for binding His-Phe, because we already knew, from prior experiments, the prominent sites for the amino acids: histidine and phenylalanine. A bit of selection trickery was required. When selecting for peptide binding, affinity for the one amino acid side chain that RNA can bind with the fewest nucleotides is usually recovered instead. This, however, can be avoided by counter-selecting against affinity for histidine or phenylalanine alone.

With such counterselection, we relatively quickly recovered at least 18 kinds of RNA sequence that bound His-Phe. Five of these were characterized carefully, and, indeed, bound both His and Phe peptide side chains in the dipeptide, with little affinity for His or Phe monomers. A notable single RNA had His-Phe affinity with weak, but measurable, affinity for His and Phe, despite counterselection.

What do these findings mean? Two amino acids in a peptide can indeed be simultaneously bound to RNA; necessarily, they are bound with only the short covalent peptide linkage between them. To this extent, DRT is supported. Interestingly, none of the newly selected His-Phe binding RNAs contained previously known, multiply-isolated His or Phe site sequences. Thus, the easiest peptide sites to find are not linear chains of individual amino acid sites. In fact, the new peptide-binding structure seems to have ca. 35% fewer nucleotides than the sum of the prior His and Phe sites. This result explains why these His-Phe sites are the more frequent result of selection. A salient question remains: can a DRT-like template join activated amino acids and emulate the peptidyl transferase reaction of the ribosome? Stay tuned.

Beyond DRT: though it was not an object of the study, the two peptide binding RNAs that were repetitively isolated had site sequences that contained no Phe coding triplets (as Phe sites themselves had not), but exhibited excess of RUG His anticodons (R=A or G), as observed in previous His sites. A major newly selected His-Phe site is shown in the Figure, with observed His anticodons in red circles. Thus, by a new route that finds a new class of RNA binding sites, His affinity is again accompanied by His coding triplets.

A coda in triplets. Amino acids of varied kinds, hydrophobic as well as polar, are definitely bound in RNA sites that are side chain-specific, as well as stereospecific. The simplest (most frequent) among these sites have significant associations with the genetic code, showing unexpected frequencies of cognate codons and anticodons. Amino acids in peptide linkages can be bound to RNA in an ordered, specific fashion, and they are also accompanied by unselected coding triplets, presenting the question of whether DRT might have been the primitive form of coded protein synthesis.

These findings support a “stereochemical” origin of the code, the quoted word suggesting a chemical relation between the bound amino acid and site triplet sequences. However, if conserved binding site triplets do not interact with amino acids but are just serving other essential site purposes, they might also have been captured when the genetic code was founded. This distinction might be used to define “direct” and “less direct” stereochemical origins. I currently favor the direct stereochemical origin because:

1. It is more consistent with the observation that, when the smallest binding sites are forced, the concentration of cognate coding triplets goes up (2).
2. It is more consistent with the observation that a new class of RNA binding sites for peptides (DRT) also shows signs of cognate triplet involvement.
3. In two informative RNA site structures, nucleotides implicated in specificity are a) in Arg codons, and b) in atomic contact with bound arginine (4, 5).

References

1. Yarus M (1988) A specific amino acid binding site composed of RNA. *Science* 240:1751-1758.
2. Yarus M, Widmann JJ, & Knight R (2009) RNA-amino acid binding: A stereochemical era for the Genetic Code. *Journal of Molecular Evolution* 69:406-429.
3. 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(3-4):217-225.
4. Yang Y, Kochoyan M, Burgstaller P, Westhof E, & Famulok F (1996) Structural Basis of Ligand Discrimination by Two Related RNA Aptamers Resolved by NMR Spectroscopy. *Science* 272:1343-1346.

5. Michel F, Hanna M, Green R, Bartel DP, & Szostak JW (1989) The guanosine binding site of the *Tetrahymena* ribozyme. *Nature* 342:391-395.

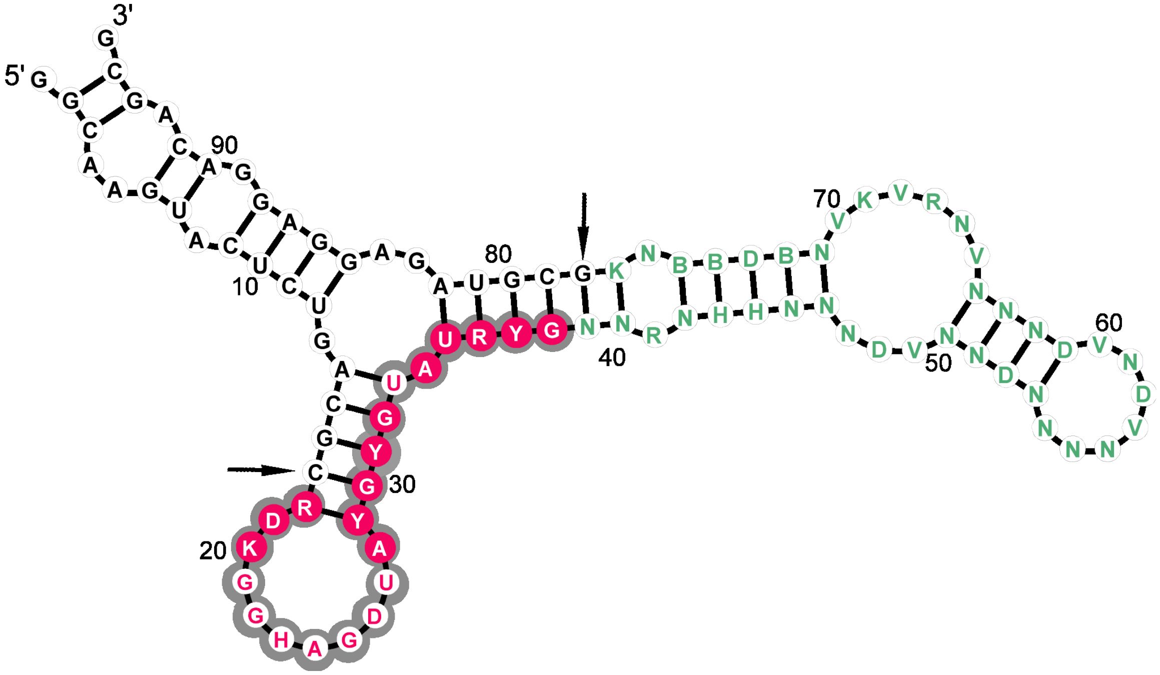

Figure: A major His-Phe peptide binding sequence, with potential RUG His anticodons marked. A secondary structure stable for multiple occurrences of the conserved site motif (circled nt) is shown; the conserved motif is threaded through this stable structure.. Variable nucleotides are written as IUPAC consensus, whose symbols are: *N* – A/C/G/U, *R* – G/A, *Y* – C/U, *M* – A/C, *K* – G/U, *W* – A/U, *S* – C/G, *B*- not A, *D* – not C, *H* – not G, *V* – not U. Arrows point to innermost primer nucleotides. Highly conserved motif nucleotides are circled in gray – potential His RUG anticodons are white letters on red circles, conserved nucleotides unrelated to his triplets are red letters on white circles. Note that some triplets are complementary to constant primer complements. Primer complements are black; initially randomized nucleotides outside the highly conserved binding site motif are green.