PERSPECTIVE

RNA–ligand chemistry: A testable source for the genetic code

MICHAEL YARUS

Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, Colorado 80309-0347, USA

ABSTRACT

In the genetic code, triplet codons and amino acids can be shown to be related by chemical principles. Such chemical regularities could be created either during the code's origin or during later evolution. One such chemical principle can now be shown experimentally. Natural or particularly selected RNA binding sites for at least three disparate amino acids (arginine, isoleucine, and tyrosine) are enriched in codons for the cognate amino acid. Currently, in 517 total nucleotides, binding sites contain 2.4-fold more codon sequences than surrounding nucleotides. The aggregate probability of this enrichment is 10^{-7} to 10^{-8} , had codons and binding site sequences been independent. Thus, at least some primordial coding assignments appear to have exploited triplets from amino acid binding sites as codons.

Keywords: amino acids; binding site; coding; origin; selection-amplification; triplets

INTRODUCTION

The RNA World hypothesis (Gilbert, 1986) makes specific predictions, as illustrated by the detection of many new RNA activities during its exploration (for review, see Yarus, 1999). For example, succession from the RNA World to modern biology likely required synthesis of large, specific proteins. Such proteins subsequently replaced many RNA functions, allowing the evolution of organisms with modern informational mechanisms. Thus translation and the genetic code were required for exit from the RNA World. Accordingly, one predicts that coding and other translational features originate in processes that predate this exit, that is, in RNA–amino acid chemistry.

This RNA World point of view recapitulates prior arguments. The origin of the genetic code has frequently been cast in terms of two alternatives. One hypothesis was championed by Carl Woese (Woese, 1967), who argued that there was stereochemical matching, that is, affinity, between amino acids and certain triplet sequences. Woese et al. (1966) write: "All in all, the conclusion that the genetic code derives basically from some sort of codon–amino acid-pairing interactions is essentially unavoidable."

The other line was taken by F.H.C. Crick (1968), who, in a well-known argument, considered that the basis of the code might be a "frozen accident," with no underlying chemical rationale. However, Crick explicitly left room for stereochemical interactions between a few amino acids and their coding nucleotides, and urged exploration of such interactions as the avenue to empirical investigation of the code: "It is therefore essential to pursue the stereochemical theory... vague models of such interactions are of little use. What is wanted is direct experimental proof that these interactions take place ... and some idea of their specificity."

The genetic code is a set of rules relating nucleotide triplets and amino acids. These relations are visibly regular. For example, all codons with a central U are cognate to amino acids with hydrophobic side chains, and codons with a central A are cognate to amino acids with polar side chains. Such chemical regularities among triplets for similar amino acids are seemingly congenial to a stereochemical origin for the coding rules.

Not only is such chemical order apparent by inspection, but it can be made more quantitative and compelling. Woese et al. (1966) measured the chromatographic mobility of amino acids on paper in varying solvents, and interpreted mobility in terms of the number of water molecules required in the solvation sphere of a mobile molecule, calling this the "polar requirement." He noted that similar codon triplet sequences were associated with amino acids of similar polar requirement. In fact,

Reprint requests to: Michael Yarus, Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, Colorado 80309-0347, USA; e-mail: yarus@stripe.colorado.edu.

polar requirement is a useful index that captures the order in the coding table better than some possible alternatives, such as hydropathy, molecular volume, or isoelectric point (Haig & Hurst, 1991). If the probability of various kinds of translational errors is taken into account, the natural code minimizes the effects of amino acid substitution on polar requirement in proteins exceptionally well. It is better than all but one in a million randomly generated codes, constituted by moving amino acids arbitrarily among the existing coding blocks (Freeland & Hurst, 1998a). A somewhat related idea, that coding order reflects the successive acquisition of related codons by biosynthetically related (and therefore chemically related) amino acids (Wong, 1988) now seems less appealing. Biosynthetic coevolution gives an account of the existing genetic code that is worse than many randomly chosen codes in which amino acids are arbitrarily assorted to the existing coding blocks (Amirnovin, 1997). Further, imposition of biosynthetic constraints does not explain the natural order of the code (Freeland & Hurst, 1998b). Thus reassignment of codons to metabolic descendants may have occurred, but is not likely to be the overall principle that ordered the code. The almost universal code therefore assigns amino acids to related triplets on the basis of a property related to polarity/hydrophobicity, well-captured by the measurement of polar requirement. In fact, given the apparent predictive value of Woese's polar requirement, a better experimental and theoretical definition of it would be very desirable.

However, it is implausible that the complex structure of the code would appear in near-final form, as emphasized by both Crick and Woese. Instead, the code would grow from a small nucleus of primordial assignments. Most codons would be used to minimize truncation of peptides by nonsense termination. However, initial assignments would be only partially specific, perhaps sets of codons to sets of amino acids. Such an initial code would evolve by more specific reassignment of subgroups of codons to particular amino acids. Successful reassignments were predicted to minimize disruption of the structures of existing proteins (Crick, 1968) or minimize the effect of translational errors by ensuring that errors inserted similar amino acids (Woese, 1967).

Remarkably, even modern organisms can shift amino acids between codons despite a complex genome encoding thousands of highly evolved proteins. The potentially deleterious effects of general amino acid substitution on protein structure at first seem daunting. However, fatal effects from genome-wide amino acid substitution can likely be averted by one or two mechanisms. In one case, genetic drift enables the loss and reappearance of codons that become very rare (Osawa & Jukes, 1989). Alternatively, ambiguous coding intermediates exist in which the old and new codon assignment coexist transiently (Yarus & Schultz, 1997) without totally inactivating any gene product. It is yet more likely

that early in the code's history, with fewer constraints and more ambiguous translation, frequent reassignment would have shaped the coding table. Constrained reassignments that minimize perturbation of protein structures (Crick, 1968), and/or minimize the impact of translational errors (Woese, 1967) could ultimately yield the familiar nearly universal code. Such reassignments might have been separate from, and ended long after, the primordial assignment process.

Thus despite the new evidence above that codons and amino acid side chains have been forced into their current relation by a chemical logic, we do not know how that logic was applied. It can be divided conceptually in any proportion between primordial assignments and the later expansion of the code by reassignment. Origin studies therefore require an independent way to recognize primordial assignments.

AMINO ACID BINDING SITES

Such an independent kind of evidence comes from RNA sites that bind amino acids. Within such binding sites ribonucleotide sequences and amino acids are necessarily physically and chemically related. The first-discovered of such sites was an L-arginine site bound within the guanosine splicing cosubstrate site of the group I intron (Yarus, 1988). Notably, the broad conservation of this interaction implies some antiquity, and the triplet sequence embracing bound G/arginine consists of five of six arginine codons only in hundreds of examples (reviewed in Yarus, 1993). This persistent association between arginine triplets and bound arginine suggested (Yarus & Christian, 1989) that the codons of the genetic code might have originated by the extraction of triplet sequences present within primordial RNA binding sites. The subsequent invention of selection-amplification allowed the systematic exploration of the ribonucleotide sequences within many new amino acid binding sites in RNAs selected for affinity for amino acid resins (Famulok & Szostak, 1992).

THE PREDICTION

If prevalent triplet sequences within ancient RNA–amino acid binding sites were later abstracted to serve as codons in the modern code, then we predict coding triplets within a subset of modern selected sites. Not all amino acid sites need have coding sequences. For one thing, only a subset of modern amino acids need have been those assigned at the code's origin (compare above). In addition, for arginine every new selection isolates new RNA binding motifs (Yarus, 1998). Therefore the population of possible RNA–amino acid binding sites can be large, effectively beyond exhaustive survey by selection-amplification. We cannot know if contemporary selections isolate the correct subset of sites. Thus, we cannot predict what fraction of a set of

modern sites will contain codons. Instead we hope that the historical selection of particular triplets for the almost universal code implies that their sequences were once frequent. Thus they could plausibly recur within modern amino acid binding sites often enough that a convincing statistical demonstration of their localization can be constructed.

Arginine codons do occur more frequently than expected in selected RNA binding sites for this amino acid (Yarus, 1998). However, arginine in particular poses an attendant statistical problem: because it has six codons, a convincing test must take full account of the fact that arginine triplets will recur frequently without any particular selection for them.

THE TEST

A statistical test with the power required was devised by Knight and Landweber (1998). It relies on comparison of the coding triplet sequences that exist within the amino acid binding site (defined by structural or chemical data) with triplet sequences in the same oligomer but in nonsite nucleotides. The question asked is whether coding triplet nucleotides exist among the site nucleotides in greater proportion than in the surrounding sequences, used as an internal control. The power of the test comes from the controlled comparison between site nucleotides and their surroundings. Surrounding sequences have come from the same randomized pool, have emerged from the same selection-amplification, and now indeed, exist in the same molecule. The test is unusually free of assumptions, in that there is no reference to abstract calculation of expected triplet frequencies. The test also seems conservative, in that a finite suite of biochemical tests will probably fail to find some nucleotides near the site, underestimating the number of coding nucleotides involved in the site (see below). In its first application, this test detected an improbable concentration of triplet nucleotides in arginine sites (Knight & Landweber, 1998).

VALIDITY OF SITE-NUCLEOTIDE DEFINITION

Such calculations rely on distinction between site nucleotides and nucleotides more distant from a bound amino acid. This distinction can sometimes be based on a three-dimensional structure. However, more usually chemical criteria must be used. For examples, sequence conservation, amino acid-evoked nucleotide protections, and modification-interference data can be gathered for many sites. I now wish to gauge the consistency of different methods for site/nonsite nucleotide classification by comparing data of different types.

For example, a set of arginine/citrulline binding oligomers for which extensive chemical data were collected (Burgstaller et al., 1995) has also been analyzed to give an NMR structure (Yang et al., 1996). Figure 1

contains, at the top, a drawing of the secondary structure of the arginine-binding oligomer. Below it is a comparison of the importance of each of the nucleotides surrounding the amino acid binding site (within the large internal loop) in the NMR structure. The structure can be compared to the criteria of conservation in different isolates, chemical accessibility on binding of arginine, and modification-interference, using binding to arginine-agarose to assess interference.

Agreement among the four types of evidence is remarkable. Nucleotides G9, G12, C13, A29, G30, G31, A33, G34, and G35 are the “first shell” of nucleotides, whose bases are immediately adjacent to bound arginine (Fig. 1). These nucleotides take roles that vary from H-bonding directly to arginine to participation in the continuity of the walls of the arginine site. In fact, every one of the first shell nucleotides in the three-dimensional structure is scored as a significant nucleotide by all three chemical criteria—by conservation, arg protection, and modification-interference. U32 is a single potentially ambiguous case; although it is adjacent to the site, the uridine base is pointed outward, taking no direct role in liganding arginine. Viewed with chemical criteria, U32 is a conserved nucleotide among initial isolates (Famulok, 1994), but is not protected nor is it a site of modification-interference with arginine binding.

The “second shell” of site nucleotides includes A8, A11, G14, C28, and U36. These nucleotides pair to support the floor of the site (11–36) or close the other end of the site, acting as the initial pair of a distal helix (14–28). They uniformly take no direct role in arginine binding. These give varied chemical signatures. For example, A8 is neither conserved nor does it interfere, but it does increase in accessibility upon arg binding. This is distinct from the first shell nucleotides, but does suggest correctly that A8 is close to bound arginine. Neither nucleotide of the 11–36 base pair which supports the floor of the site is conserved, but the purine A11 is slightly protected and interferes. This suggests that stacking with this pair contributes to site function, but the particular paired nucleotides playing this role can vary. Thus each second-shell nucleotide for which data are available gives a partial chemical signal, which depends in a frequently intelligible way on its particular role in the NMR-defined active structure.

Finally, one stacking interaction and/or phosphodiester further away from the bound amino acid are A7, G10, C37, and A38 (cf. Fig. 1). These “third shell” nucleotides are extremely varied in their chemical signature. A7 and A38 form a tertiary pair that is neither conserved or protected, nor does it interfere. G10–C37, the second level pair beneath the site floor, has an essential purine (G10) that is conserved and protected and it interferes, as well as a notably ambiguous pyrimidine partner (C37), which is conserved, but is neither protected nor does it interfere. This may again

FIGURE 1. Role of the nucleotides in an arginine site, assessed four ways. The top panel shows the secondary structure of the RNA termed 44.Arg 11 by Famulok (1994). Chemical protection data in the bottom panel are from Burgstaller et al. (1995). Comments on the three-dimensional structure are derived from the published NMR structure (Yang et al., 1996), as well as from inspection of the deposited structures 1KOC and 1KOD in the Protein Data Base. Two arginine codons created when arginine-binding was selected are shown in bold, and the 3 nt changed in this selection (starting with a citrulline-binding RNA) are underlined.

indicate a significant purine-stacking buttress beneath the binding site.

More distal nucleotides in the "fourth shell" and beyond, always in helical base pairs in this three-dimensional structure, are not conserved and not picked up by any other chemical criterion.

To summarize: chemical indices for essential function fall off sharply outside the first shell of binding-site nucleotides. Such data unanimously indicate a first-shell nucleotide whose base (the usual focus of both conservation and modification data) has a role in the site. Put another way, the three chemical indicators, even taken individually, would each have been a reasonably reliable guide to the nucleotides nearest the amino acid. Second- and third-shell nucleotides are

distinguishable by chemical criteria from first-shell nucleotides, because they give partial chemical signals whose nature can sometimes be rationalized by the three-dimensional structure. Nucleotides at this proximity (e.g., U36) may occasionally be undetectable. More distant nucleotides give no chemical signals at all.

Thus with nucleotide conservation, protection, and interference at hand, we expect to find essentially all first-shell nucleotides, most second-shell nucleotides, and some third-shell nucleotides. Although such data is notably ambiguous with regard to specific structural role, the probability of detection of a site nucleotide by chemical means is high, and it may even be possible to partially distinguish nucleotides by proximity to the ligand. Therefore the chemical classification of nucleo-

tides as site or nonsite for statistical tests seems well justified.

RESULTS OF THE TEST FOR CODING TRIPLET DISTRIBUTION

In Figure 2 are all the presently informative amino acid sites, meaning those for which chemical data are available (L-valine is excluded for this reason; Majerfeld & Yarus, 1994). In particular, these include the recent tyrosine-binding RNAs (Mannironi et al., 2000) whose description provoked this re-analysis. These tyrosine sites are derived from an initial set of dopamine-binding RNAs (Mannironi et al., 1997). In Figure 2, sites arising from separate selections appear in separate panels. The initially randomized nucleotides that are the basis of the analysis are capitalized, and lower case is used for initially fixed sequences, appended for experimental convenience during amplification. Because fixed sequences are not results of selection, they are not included in the analysis. However, they are shown where their explicit presence clarifies the structure of the binding site. Coding triplets in initially randomized regions are in colored bold face. Chemical data (such as protections and interferences) that may identify the subset of binding-site nucleotides are summarized as symbols above or below the implicated nucleotide or internucleotide linkage.

The notable finding is that the distribution of coding triplets favors site rather than nonsite nucleotides. Frequently this unusual distribution is evident without a statistical test. For example, in the arginine site in Figure 2B or the uppermost tyrosine site in Figure 2F, clustering of codons in and near the binding site leaps to the eye. In fact, for each of these two RNAs, analysis bears out this intuitive view. Triplet nucleotides appear within these sites with a probability of less than 0.02 (arg) or 0.001 (tyr), had triplet nucleotides been placed without relation to the amino acid binding sites. One cannot base a valid statistical argument on sequences chosen to show a property (rather, sequences must be sampled without reference to the tested property). However, the calculation indicates that even single RNAs in this sample, were they the only data, would affirm an improbable concentration of triplets among binding-site nucleotides.

Figure 3 presents the complete nucleotide census for arginine, isoleucine, and tyrosine oligomers as 2×2 matrices that are useful for the statistical test, but which also allow an overall appreciation of the distribution of nucleotides. Nucleotides are classified as within and outside the site, and as members of codons or other. The idea tested is whether these classifications are independent. The global and tyrosine distributions are split into two tables because Mannironi et al. (2000) suggest that tyrosine sites had an excess of UAN codons (modern UAY tyr plus UAA/G stop). The analysis

has been done for this extended codon group as well as for UAY (modern tyrosine triplets) alone.

Remarkably, sites of every specificity show an excess of triplet nucleotides within the site nucleotides. For arginine, 68% of site nucleotides (36/53) are members of coding triplets versus 25% triplets in outside, nonsite nucleotides. For isoleucine, 25% of site nucleotides (3/12) are in triplets versus 16% outside. Tyrosine sites contain 26% UAN triplets (22/84) versus 11% outside. Alternatively, tyrosine sites have 14% UAY triplets versus 4.6% outside. In all sites currently characterized, coding nucleotides are 41% within sites versus 17% outside (or including UAY only, 34% in sites versus 14% outside). Thus triplet nucleotides are about 2.4-fold more frequent within than outside RNA–amino acid binding sites.

Not only is the disproportionate presence of triplets within amino acid sites consistent, it is highly unlikely to have occurred by sampling alone. For isoleucine, for which few site nucleotides are available for analysis, the distribution of codons into the site is observable but not statistically significant on its own. However, arginine triplets, or independently, tyrosine triplets, occur improbably frequently among the nucleotides of their amino acid's sites. The most significant case can be made by combining all available data, as at the bottom of Figure 3. There are 517 nt from 9 RNAs obtained in four different laboratories. The probability of the observed excess of triplet nucleotides among binding-site nucleotides in arginine, isoleucine, and tyrosine sites taken together is on the order of 10^{-7} to 10^{-8} , had the coding triplets been placed without regard to proximity to the site.

COMMENTS ON THE DISTRIBUTIONS

Thus a variety of selected amino acid-binding oligomers, comprising all with reported chemical data for the ligand site, overrepresent cognate codons among their binding-site nucleotides. These RNAs (Fig. 2) have usually been singled out for chemical analysis because they were the majority sequences among a greater number selected. In all cases they were originally analyzed without regard to possible codon content, within sites or elsewhere.

The RNAs (Fig. 2) are obtained by diverse techniques, and a conclusion derived from them draws some strength from their diversity. For example, purines are overrepresented in RNA binding sites for varied ligands (Knight & Landweber, 1998). However, the unusual distribution of coding nucleotides is observed with triplets in which purines are in the majority (CGN, AGR arginine) as well as for triplets composed mostly of pyrimidines (UAY tyrosine). The amino acids themselves are disparate, embracing amino acids with the most intensely polar side-chain grouping (the guanidinium of arginine) as well as both aliphatic and aromatic hydro-

FIGURE 2. A menagerie of selected amino acid binding sites. Each panel represents an independent selection. Randomized nucleotides are uppercase, initially fixed sequences (where shown) are lower case. Coding triplets are bold and green. Superscript or subscript symbols represent the outcome of chemical tests of each nucleotide's role in binding the amino acid. The selection of chemical tests is noted by abbreviations in each panel; int: interference, prot: protection, nuc: nuclease, sens: sensitization, cons: conservation. In some cases, published data have been reevaluated to define a set of site nucleotides (I hope) more consistent across experiments. Original selections can be found as follows: **A:** Connell & Yarus (1994); **B:** Famulok (1994)—RNA 44Arg11; **C:** Tao & Frankel (1996)—RNAs 2 and 16; **D:** Geiger et al. (1996)—RNA ag.06; **E:** Majerfeld & Yarus (1998)—RNA IL-19; **F:** Mannironi et al. (2000)—RNAs tyr1, 2, and 3. This list excludes selected RNA sites that do not function with free amino acids or for which chemical data is too limited. If needed, these are reviewed in Yarus (1998). (*Figure continues on facing page.*)

FIGURE 2. (Continued.)

Summary - arginine sites			
	Nt in triplets	Nt not in triplets	TOT Nt within or outside site
Nt within sites	36	17	53
Nt outside sites	38	105	143
TOT triplet/non-triplet	74	122	196 total Nt
$\chi^2 = 26.6, p = 1.3 \times 10^{-7}$			
Summary - isoleucine site			
Nt within site	3	9	12
Nt outside site	6	32	38
TOT triplet/non-triplet	9	41	50 total Nt
$\chi^2 = 0.086, p = 0.38$			
Summary - tyrosine sites with UAN codons			
Nt within sites	22	62	84
Nt outside sites	20	167	187
TOT triplet/non-triplet	42	229	271 total Nt
$\chi^2 = 9.48, p = 0.001$			
Summary - tyrosine sites with UAY codons			
Nt within sites	12	72	84
Nt outside sites	9	178	187
TOT triplet/non-triplet	21	250	271 total Nt
$\chi^2 = 6.93, p = 0.004$			
Global summary of all amino acid sites (using UAN for tyrosine)			
Nt within sites	61	88	149
Nt outside sites	64	304	368
TOT triplet/non-triplet	125	392	517 total Nt
$\chi^2 = 30.8, p = 1.4 \times 10^{-8}$			
Global summary, all sites (using UAY for tyrosine)			
Nt within sites	51	98	149
Nt outside sites	53	315	368
TOT triplet/non-triplet	104	413	517 total Nt
$\chi^2 = 24.7, p = 3.3 \times 10^{-7}$			

FIGURE 3. Distribution of the coding nucleotides in amino acid-binding RNAs. Nucleotides (Nt) are counted as being within a binding site if they are protected by the ligand, or if their modification interferes with later binding. Such tests for base modification implicate only the nucleotide modified. However ENU effects or nuclease effects, because they apply to an internucleotide linkage, are taken to implicate both flanking nucleotides in the binding site. Only experimentally randomized nucleotides are included in the tests, because the identities of these nucleotides alone are determined by the selection for amino acid affinity. Although fixed sequences (lower case) are sometimes shown in Figure 2, they are not counted in any category of the tables in this figure, even when they may be a part of the binding site. To the right and below each 2×2 matrix is its chi-squared, calculated using Yates' correction for continuity. Probabilities are for a 1-tailed test at 1 degree of freedom, testing for the independence of being a coding triplet nucleotide from being identified as a binding site nucleotide. This is the standard statistical test for independence of classifications. In this application, it gives the same probabilities as the G test used by Knight & Landweber (1998).

phobic side chains (isoleucine and tyrosine). Nevertheless, the same tendency is observed in these distributions. Triplet excess has now been observed for sites derived directly from random pools (arginine, isoleucine), as well as for those derived by remutagenesis and reselection for a new ligand affinity (some arginine, tyrosine). Disparate routes to similar conclusions suggest the result is robust, and will be easily observable when an expanded selection of sites is available for analysis.

Furthermore, observation of elevated triplet nucleotides for three of three arbitrarily selected amino acid sites suggests that such behavior is relatively frequent. Given that three of three amino acids are positive, the true proportion of amino acids with codons in their binding sites is between 0.29 and 1.0, with 95% confidence

(calculated as by Clopper & Pearson, 1934). Therefore one can plausibly expect 6–20 amino acids with excess binding-site codons. There may be some bias in this sample, because only successful selections (and therefore amino acids that interact with RNA) can be considered. Nevertheless, a substantial part of the chemical order of the code is likely a consequence of the initial codon assignment process rather than later evolution. Therefore a stereochemical hypothesis (Woese, 1967) in which codons are abstracted from larger RNA structures that form amino acid binding sites is strongly supported by present data.

Finally, the clue supplied by the natural binding of arginine within arginine codons in the group I active center (Yarus, 1988, 1998) is now reinforced by selected sequences (Fig. 2). The possibility that the

group I–arginine association is a true molecular fossil from the code's initiation is thereby strengthened. Previous evidence suggests that the group I intron predates the origin of the plant chloroplast from an invading cyanobacterium (reviewed in Paquin et al., 1997). Coding-fossil status would make group I RNA, or at least its splicing cofactor site, much more ancient, placing it squarely within the RNA World.

Acceptance of the hypothesis that codon-enriched RNA structures initially bound the amino acids radically divides prior discussion about the origin of the genetic code. For example, anticodons then have nothing to do directly with amino acids (compare Lacey et al., 1985). Thus anticodon triplets and tRNA-like structures are by implication later intercalations between codons and amino acids. For example, it has been suggested that tRNAs were selected later to standardize the cycle of peptide bond-forming reactions (Yarus, 1998).

FOR THE FUTURE

It would clearly be of great interest to have a broader selection of sites, obtained and analyzed by a suite of consistent techniques. This has to do with the desirability of further reducing the probability of the null hypothesis; that is, that coding triplets and amino acid binding sites in RNA have nothing to do with each other. However, this notion is already quite improbable (Fig. 3); the signal (triplets within sites) seems much stronger than the minimal prediction of the direct RNA template theory (above and Yarus, 1998).

Therefore the apparently good signal/noise ratio in these early results raises the hope that subtler aspects of the code's history may be revealed when selected RNA binding sites are broadly studied. To illustrate the flavor of such argument, consider the preliminary but growing data on aromatic amino acids. As noted above, Mannironi et al. (2000) emphasize the excess of UAN in selected tyrosine binding sites. Perhaps all UAN codons were assigned to tyrosine initially, rather than being partially nonsense. In fact, the tyrosine binding sites of Mannironi et al. (2000) also bind tryptophan and, more weakly, phenylalanine. Previous and independently isolated phenylalanine ribonucleotide binding sites have a related property, also binding tryptophan, though this affinity was not selected (Zinnen & Yarus, 1995). In addition, a highly specific ribozyme that synthesizes phe-RNA from adenylate discriminates against other amino acids by more than four orders of magnitude. This ribozyme is more specific than is the modern protein, yeast PheRS. Yet, this highly specific catalyst for phe-RNA synthesis still accepts tyrosine at a slightly reduced rate (Illangasekare & Yarus, 1999). These observations would likely be clarified by a fully characterized (cf. Fig 2) set of phenylalanine and tryptophan sites for comparison with tyrosine. Nevertheless, they already suggest that readily selected RNA sites for ar-

omatic amino acid side chains will employ a relatively nonspecific principle like stacking, and that tyrosine, phenylalanine, and tryptophan accordingly may be relatively poorly distinguished. An inclusive hypothesis therefore is that UAN initially meant "aromatic," a type of ambiguity that has long been suspected for the early genetic code (Woese, 1967; Crick, 1968). This contrasts strongly with polar side chains that are easily distinguished by RNA. For side chains such as arginine, coding may have been specific from the code's origin (Yarus, 1993).

However plausible or implausible such speculations appear in the light of more complete data, the immediate future of the code's origin appears to be in the hands of experimentalists, where Woese asserted it and Crick wished it more than three decades ago.

ACKNOWLEDGMENTS

Thanks to Martin Horvath for help with molecular graphics and to Rob Knight, Leslie Leinwand, Norm Pace, and Bob Thompson for helpful comments on a draft manuscript, though the views expressed should be attributed only to the author. MY's laboratory was supported by National Institutes of Health research grants GM30881 and GM48080 during work described here.

REFERENCES

- Amirnovin R. 1997. An analysis of the metabolic theory of the origin of the genetic code. *J Mol Evol* 44:473–476.
- Burgstaller P, Kochoyan M, Famulok M. 1995. Structural probing and damage selection of citrulline- and arginine-specific RNA aptamers identify base positions required for binding. *Nucleic Acids Res* 23:4769–4776.
- Clopper CJ, Pearson ES. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26:404–413.
- Connell G, Yarus M. 1994. RNAs with dual specificity and dual RNAs with similar specificity. *Science* 264:1137–1141.
- Crick FHC. 1968. The origin of the genetic code. *J Mol Biol* 38:367–379.
- Famulok M. 1994. Molecular recognition of amino acids by RNA-aptamers: An L-citrulline binding RNA motif and its evolution into an L-arginine binder. *J Am Chem Soc* 116:1698–1706.
- Famulok M, Szostak JW. 1992. Stereospecific recognition of tryptophan-agarose by in vitro selected RNA. *J Am Chem Soc* 114:3990–3991.
- Freeland SJ, Hurst LD. 1998a. The genetic code is one in a million. *J Mol Evol* 47:238–248.
- Freeland SJ, Hurst LD. 1998b. Load minimization of the genetic code: History does not explain the pattern. *Proc R Soc Lond B Biol Sci* 265:2111–2119.
- Geiger A, Burgstaller P, von der Eltz H, Roeder A, Famulok M. 1996. RNA aptamers that bind L-arginine with sub-micromolar dissociation constants and high enantioselectivity. *Nucleic Acids Res* 24:1029–1036.
- Gilbert W. 1986. The RNA World. *Nature* 319:618.
- Haig D, Hurst LD. 1991. A quantitative measure of error minimization in the genetic code. *J Mol Evol* 33:412–417.
- Illangasekare M, Yarus M. 1999. Specific, rapid synthesis of phe-RNA by RNA. *Proc Natl Acad Sci USA* 96:5470–5475.
- Knight RD, Landweber LF. 1998. Rhyme or reason: RNA-arginine interactions and the genetic code. *Chem Biol* 5:R215–R220.
- Lacey JC Jr, Hall LM, Mullins DW Jr. 1985. Rationalization of some genetic anticodon assignments. *Orig Life Evol Biosph* 16:69–79.
- Majerfeld I, Yarus M. 1994. An RNA pocket for an aliphatic hydrophobe. *Nature Struct Biol* 1:287–292.

- Majerfeld I, Yarus M. 1998. Isoleucine: RNA sites with associated coding sequences. *RNA* 4:471–478.
- Mannironi C, DiNardo A, Fruscoloni P, Tocchini-Valentini GP. 1997. In vitro selection of dopamine RNA ligands. *Biochemistry* 36:9726–9734.
- Mannironi C, Scherch C, Fruscoloni P, Tocchini-Valentini GP. 2000. Molecular recognition of amino acids by RNA aptamers: The evolution into a L-tyrosine binder of a dopamine binding motif. *RNA* 6:520–527.
- Osawa S, Jukes TH. 1989. Codon reassignment (codon capture) in evolution. *J Mol Evol* 28:271–278.
- Paquin B, Kathe SD, Nierzwicki-Bauer SA, Shub DA. 1997. Origin and evolution of group I introns in cyanobacterial tRNA genes. *J Bacteriol* 179:6798–6806.
- Tao J, Frankel AD. 1996. Arginine-binding RNAs resembling TAR identified by in vitro selection. *Biochemistry* 35:2229–2238.
- Woese CR. 1967. *The genetic code: The molecular basis for gene expression*. Ch. 6–7. New York: Harper & Row. pp 156–160.
- Woese CR, Dugre DH, Saxinger WC, Dugre SA. 1966. The molecular basis for the genetic code. *Proc Natl Acad Sci USA* 55:966–971.
- Wong JT. 1988. Evolution of the genetic code. *Microbiol* 5:174–181.
- Yang Y, Kochoyan M, Burgstaller P, Westhof E, Famulok M. 1996. Structural basis of ligand discrimination by two related RNA aptamers resolved by NMR spectroscopy. *Science* 272:1343–1347.
- Yarus M. 1988. A specific amino acid binding site composed of RNA. *Science* 240:1751–1758.
- Yarus M. 1993. An RNA–amino acid affinity. In: Gesteland RF, Atkins JF, eds. *The RNA world*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press. pp 205–217.
- 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. 1999. Boundaries for an RNA world. *Curr Opin Chem Biol* 3:260–267.
- Yarus M, Christian EL. 1989. Genetic code origins. *Nature* 342:349–350.
- Yarus M. & Schultz DW. 1997. Further comments on codon reassignment. Response. *J Mol Evol* 45:3–6.
- Zinnen S, Yarus M. 1995. An RNA pocket for the planar aromatic side chains of phenylalanine and tryptophan. *Nucleic Acids Symp Ser* 33:148–151.