J Mol Evol (2009) 69:406–429
DOI 10.1007/s00239-009-9270-1

RNA–Amino Acid Binding: A Stereochemical Era for the Genetic Code

Michael Yarus · Jeremy Joseph Widmann · Rob Knight

Received: 13 May 2009 / Accepted: 28 July 2009 / Published online: 1 October 2009

Abstract By combining crystallographic and NMR structural data for RNA-bound amino acids within riboswitches, aptamers, and RNPs, chemical principles governing specific RNA interaction with amino acids can be deduced. Such principles, which we summarize in a "polar profile", are useful in explaining newly selected specific RNA binding sites for free amino acids bearing varied side chains charged, neutral polar, aliphatic, and aromatic. Such amino acid sites can be queried for parallels to the genetic code. Using recent sequences for 337 independent binding sites directed to 8 amino acids and containing 18,551 nucleotides in all, we show a highly robust connection between amino acids and cognate coding triplets within their RNA binding sites. The apparent probability (P) that cognate triplets around these sites are unrelated to binding sites is $\cong 5.3 \times 10^{-45}$ for codons overall, and $P \cong 2.1 \times 10^{-46}$ for cognate anticodons. Therefore, some triplets are unequivocally localized near their present amino acids. Accordingly, there was likely a stereochemical era during evolution of the genetic code, relying on chemical interactions between amino acids and the tertiary structures of RNA binding sites. Use of cognate coding triplets in RNA binding sites is nevertheless sparse, with only 21% of possible triplets appearing. Reasoning from such broad recurrent trends in our results, a majority (approximately 75%) of modern amino acids entered the code in this stereochemical era; nevertheless, a minority (approximately 21%) of modern codons and anticodons were assigned via RNA binding sites. A Direct RNA Template scheme embodying a credible early history for coded peptide synthesis is readily constructed based on these observations.

Keywords Genetic code · Stereochemical · Origin · Amino acid · Binding site · RNA

M. Yarus (✉) · J. J. Widmann · R. Knight
Departments of MCD Biology and Chemistry/Biochemistry, University of Colorado, Boulder, CO 80309-0347, USA
e-mail: yarus@stripe.colorado.edu

J. J. Widmann
e-mail: Jeremy.Widmann@Colorado.EDU

Introduction

> I am particularly struck by the difficulty of getting [the genetic code] started unless there is some basis in the specificity of interaction between nucleic acids and amino acids or polypeptide to build upon. (Woese 1967)

> Nonetheless, it is clear that at some early stage in the evolution of life the direct association of amino acids with polynucleotides, which was later to evolve into the genetic code, must have begun. (Orgel 1968)

Part I: The Observed Mechanism of RNA–Amino Acid Interaction

Just above, Carl Woese and Leslie Orgel, writing at the dawn of molecular biology and coding, suppose that chemical interactions between nucleotide sequences and amino acids are an indispensable basis for the genetic code. It is the conclusion of the present narrative that such interactions are easily demonstrated, utilize plausible, simple chemistry, and can indeed be shown to echo part of the genetic code.

Part I relies on recent structural work. Three-dimensional information that includes RNA-bound amino acids at high resolution is now well known, such as the

methionines in three different riboswitches specific for *S*-adenosyl methionine (Gilbert et al. 2008; Lu et al. 2008; Montange and Batey 2006). Moreover, structures for the aptamer domain for the lysine riboswitch (Garst et al. 2008; Serganov et al. 2008) and an aptamer for citrulline and arginine (Yang et al. 1996), and a natural arginine binding site (Pugilisi et al. 1993) can also be consulted. Comparison and moderate extrapolation from these structures suggest that it will be possible for RNA sites to exist that bind most of the 20 major amino acids, though the abundances of RNA structures containing sites will likely vary, as will their affinity and discrimination.

The first order of business is a rationale for binding based on characterized amino acid sites. This is both logically and historically important, because theories for the origin of the genetic code based on affinity were discouraged by the supposed lack of such sites in early experiments, such as Paul Doty's early experiments on rRNA. This skepticism about RNA–amino acid affinity persisted at least until an amino acid binding site was located in the group I active center (Yarus 1988; Yarus and Christian 1989).

In recent experiments, RNA has proven to have unforeseen chemical versatility (Chen et al. 2007). In particular, RNA is able to specifically bind varied ligands, both fitting them to preformed binding sites and more commonly, adaptively surrounding them (Hermann and Patel 2000). Conformational change that surrounds a ligand is important because it means that the bound state contains not only the four nucleotides, but also incorporates the chemical capabilities of a non-nucleotide ligand. Thus a limited chemical environment, composed of only four related ribonucleotide monomers, is enriched by the binding reaction, which adds new bonds and new structural opportunities. It is this enveloping conformational adaptation, of course, that is exploited for the throwing of regulatory riboswitches. For many RNA ligands, sites of varied complexity exist, with more complex sites capable of generally tighter binding (Carothers et al. 2004).

The Polar Profile

We summarize known structures by saying that RNA looks at polar and nearby profile elements of a bound amino acid. That is, it fixes the amino acid primarily by polar forces directed at charged or partially charged atoms and groups, and then inspects the nearby neutral profile to make sure it has the correct shape and extension. In this way, RNA can assess even partially aliphatic ligands.

Polar elements are easily detected by directing hydrogen bonds or ionic or partial charges to them. Such groups are plentiful in RNA. Therefore, it is straightforward for RNA to bind an amino acid via its polar features. An important initial implication is that all free amino acids may be RNA bound, because the α-amino and α-carboxyl are always present, supplying good complements to the hydrogen bonding donors and acceptors, for example, at the peripheries of bases, base pairs, and base triples. Even if the carboxyl is uncharged due to esterification by an activating leaving group (as in the adenylates or ribose esters which are presently universal translation substrates), the ester will offer its lone pairs of electrons as a hydrogen bond acceptor. The shape and extension (profile) of the side chain can be measured starting from these common polar points.

For example, in the type I ("SAM box") riboswitch, the replacement (by hydrogen) of either the α-amino or α-carboxyl of SAM methionine reduces its apparent K_D by more than 33,000-fold (Lim et al. 2006). This is well explained by crystallography of the type I riboswitch SAM aptamer. The structure reveals a base triple combining helical and joining nucleotides to precisely engage both α-carbon substituents of the SAM methionine via a reticulum of hydrogen bonds (Montange and Batey 2006). Similar α-carbon–nucleobase interactions, particularly via G nucleotides, are frequent (similar interactions in the lysine riboswitch are discussed below) (Fig. 1).

Or alternatively, if the amino acid has a polar side chain, the focus of the binding site can be the polar group of the side chain. This alternative case, a single-ended site focused on the side chain, is the method of the binding site in a citrulline–arginine aptamer (Yang et al. 1996), which we will discuss further below.

An important addition to the above 'primary features', from which a profile can be measured, is that aromatic rings may be counted among such features. First of all, such separation of π-electrons from nuclear charges, even if symmetrical, produces quadrupole moments. These enable aromatic rings to interact as polar elements with approaching cations, forming bonds that are as strong as

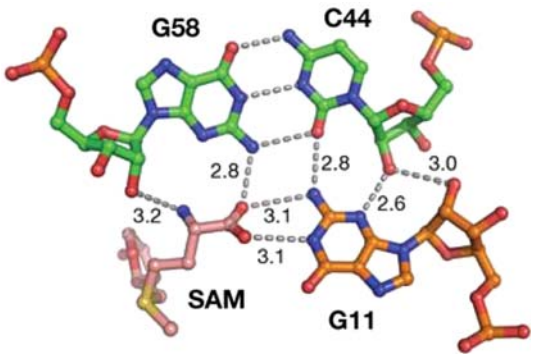


Fig. 1 Interaction with α-carbon groups: a GCG nucleotide triple within the binding site of the SAM I aptamer (Montange and Batey 2006) forms four hydrogen bonds to the α-amino (*blue* nitrogen) and the α-carboxyl (*red* oxygens) of SAM methionine

other normal secondary interactions in water (Dougherty 1996). In other words, aromatic rings are polar elements, without metaphor or approximation, and as one result, form π-cation bonds (Fig. 2).

Secondly, the layering of aromatic amino acid side chains (phe, tyr, trp, his, arg) on nucleobases, via the complex of entropic, polar, and fluctuation forces known as the stacking interaction, has been known as a primary mode of amino acid–nucleic acid interaction since the earliest nucleoprotein structures became available (Nagai 1996). A recent and striking specific example (Fig. 3) comes from the Puf proteins or pumilio homology domains, which are a repeated α-helical structure in which each helical repeat supplies an amino acid side chain to stack on and sandwich successive bound RNA bases (Wang et al. 2002). Thus by either of the above two means, an aromatic amino acid side chain may also be a group that strongly localizes an amino acid so that its adjacent profile may be determined.

When amino acid side chains contain charged or other intensely polar sites, a full repertoire of RNA groups is available to interact with side chain atoms. A salient example is the aptamer of the lysine riboswitch (Garst et al. 2008).

With its α-amino and α-carboxyl fixed by interaction with multiple G's, somewhat in the manner of the SAM I aptamer above, the charged lysine side chain amino group is forked by a pair of backbone hydrogen bonds to ribose oxygen and non-bridging phosphate oxygens (Fig. 4). These interactions are probably important, because

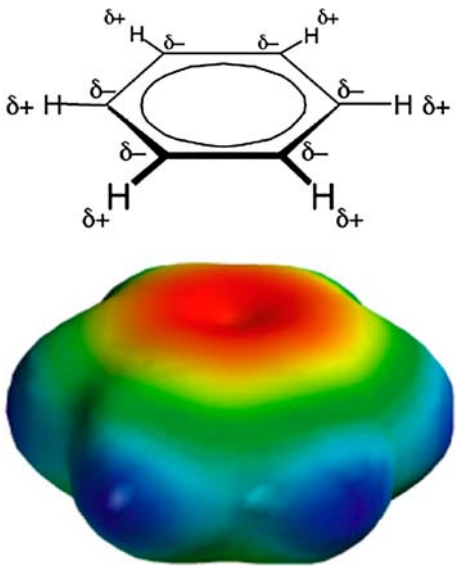


Fig. 2 The benzene quadrupole (Dougherty 2007) offers possibilities for interaction. The six C–H dipoles (*above*) create the electrostatic field shown (*below*). *Red* is negative, *blue* positive. The potential for binding a cation to the top of the ring appears clearly in the lower image

incorporation of lysine into a peptide cuts off interaction (Sudarsan et al. 2003). Thus both ends of bound lysine, and all its polar sites, are fixed by a significant confluence of multiple directional bonds to RNA. Recurrence of interactions between α-carbon substituents and an array of G-containing base pairs in the lysine riboswitch confirms that such interactions are probable, independent of mooring of the amino acid to the RNA site via an adenosine residue, as occurs in SAM aptamer structures (e.g., in Fig. 1 above).

Of course, even the aliphatic sections of side chains can interact with nucleobases by van der Waals and hydrophobic (entropic) interactions. The aliphatic inner side chain of lysine is sandwiched between the base planes of nearby purines (Garst et al. 2008; Serganov et al. 2008)—thus extended; its length can be measured by contacts with the polar groups at both ends. This suffices to distinguish lysine from similar amino acids like ornithine, whose side chain is one methylene shorter, consequently binding markedly more weakly than lysine (Sudarsan et al. 2003). However, these aliphatic/purine base plane interactions are relatively loose and non-specific. For example, the lysine side chain is hooked, not quite completely extended in fitting its site (Fig. 4). Further, the site has been shown to tolerate variation and bulky, even polar, substitutions mid-side chain, as long as the side chain remains about the right length (Blount et al. 2007). This observed tolerance for modification of the aliphatic part of the side chain contrasts with the focused specificity for terminal polar groups. Side chain length measurement within the lysine riboswitch illustrates the simplest way that an RNA site can measure a non-polar profile near polar features.

SAM riboswitches exhibit a more sophisticated style of non-polar profile determination, specifically pointed to the atom in the methionine side chain where the essential metabolic distinction must be made. *S*-adenosyl methionine (SAM) transfers its methyl group to another molecule and becomes *S*-adenosyl homocysteine (SAH). Methylation of another molecule from SAM breaks the third covalent bond between the transferred methyl and methionine side chain sulfur. Transfer accordingly removes the sulfur's positive charge to yield the neutral sulfoether of SAH. Thus genes responsive to SAM (the substrate for methylation) usually should not respond to SAH (substrate for methylation depleted). In fact, there are a different set of SAH riboswitches to stimulate recycling of SAH to SAM (Wang et al. 2008). The difference between SAM and SAH, charge and methyl, is therefore a crucial one (Fig. 5).

Therefore, it is interesting that the three SAM riboswitches whose structure is known make the SAM/SAH distinction in similar ways. All three focus on the change in polarity, using straightforward electrostatic bonds. The SAM I site points the partial negative charges of two U ring O2 carbonyls at the charged sulfur atom of SAM

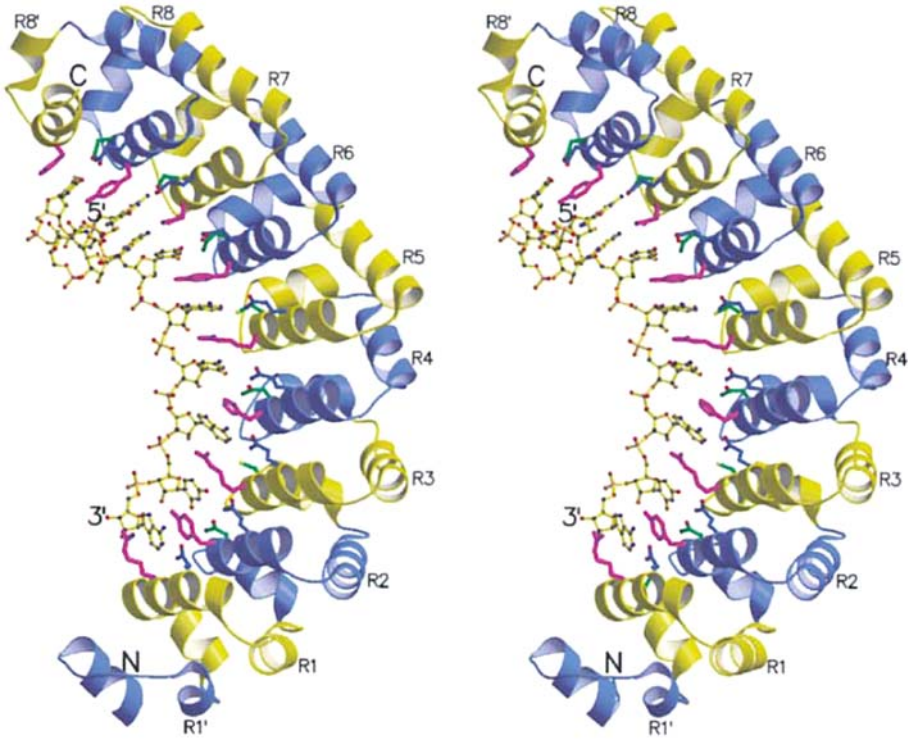


Fig. 3 Stereo pair of nucleobase-amino acid stacking within the human pumilio-homology domain bound to RNA (Wang et al. 2002). The *magenta* projections from the arc of α-helical repeats on the right are stacked amino acid side chains from position 13 of each pumilio-homology domain repeat. These amino acid side chains sandwich the nucleobases of the stick-and-ball model of bound RNA at the left

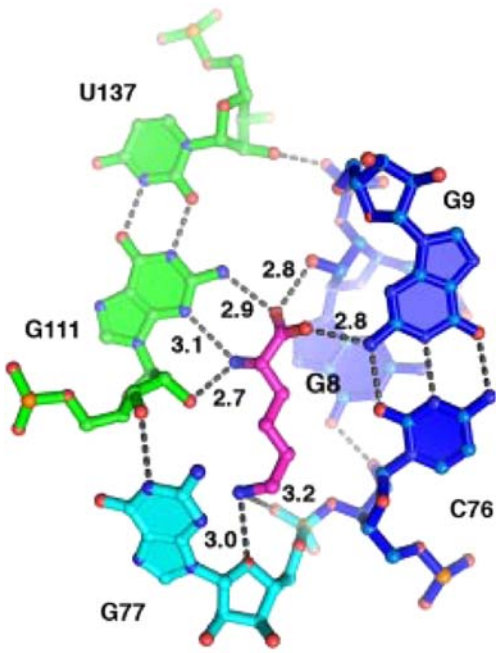


Fig. 4 Lysine bound within the lysine riboswitch (Garst et al. 2008) shows a double-ended RNA interaction. The amino acid (*magenta*) is bound via hydrogen bonds to the G's of a GC (blue) and a GU (*green*) pair. The side chain ε-amino (*bottom*) is centered by hydrogen bonds to ribose O4 and a non-bridging phosphate oxygen of a neighboring nucleotide (G77). The site requires a particular spacing between the α-carbon and ε-amino side chain groups by spanning interactions to sequential nucleotides

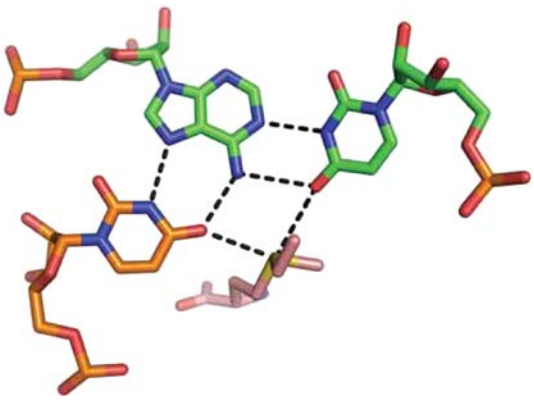


Fig. 5 RNA interaction with methylated, charged side chain sulfur–charged methionine sulfur (*pink*) within the SAM II site is recognized deep in a major groove by two U's of an AUU base triple (Gilbert et al. 2008)

(Montange and Batey 2006); the result is about an 80-fold preference for SAM (Lim et al. 2006). SAM II adopts a related strategy, using O2 and O4 of two U's from an AUU triple (Gilbert et al. 2008) near the sulfur. The SAM III strategy is parallel to these (Lu et al. 2008), but produces a >100-fold preference for SAM by directing a U O4 and a 2′ O atom from an adjacent nucleotide at the charged sulfur. This description should not be thought of as exhaustive; for example, SAM I and SAM III use a *syn*

conformation for the A; SAM II uses a more usual *anti* adenine base. However, this makes it the more remarkable that recognition of the S^+ atom is so similar in more than three cases (because of multiple molecules in asymmetric units).

Despite these similarities, the overall treatment of the sulfur substituent differs greatly between the three sites. SAM I has the methyl group pointing along the broad minor groove of a helix (Montange and Batey 2006), added bulk at this position makes little difference as long as the charge is maintained (Lim et al. 2006). SAM III points the methyl away from the site, into solvent, and makes little distinction between methyl- and ethyl-sulfonium ion (Lu et al. 2008). In fact, SAM III does not detectably interact with methionine beyond the sulfur in any sense, because no electron density is detected there.

The SAM II site (Fig. 6) uniquely makes dramatic distinctions between sulfur substituents, rejecting all alternatives to SAM by more than 1000-fold (Lim et al. 2006). This is probably because the methionine section of the SAM II site extends along the deep narrow major groove of an RNA helix, and methionine sulfur is hindered further by the third strand of a pseudoknot triplex (Gilbert et al. 2008). Thus it is clear that RNA can impose definite constraints on the size of the non-polar region near a strongly recognized polar feature, even when there is no more distal polar feature that can be strongly bound.

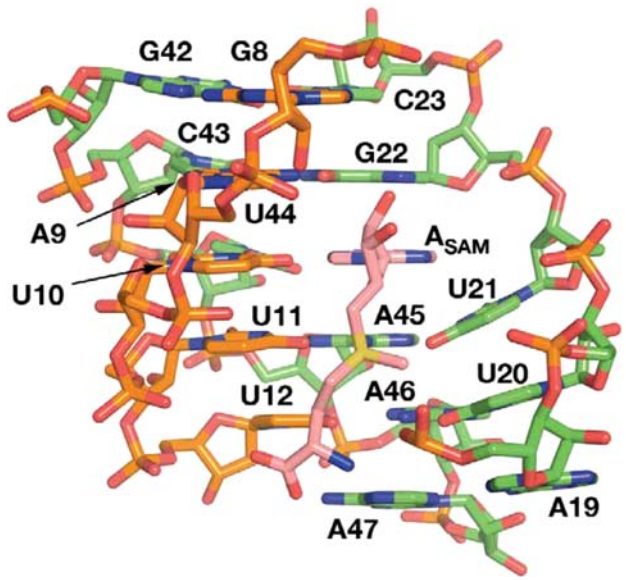


Fig. 6 SAM (*pink*) in the SAM II riboswitch site; a restrictive steric interaction (Gilbert et al. 2008)—SAM adenosine is at the top, and the α-carbon of methionine at the bottom. The snug major groove channel for the amino acid within the green and orange helical triplex is apparent

Summary of a Polar Profile, Based on Met–, Lys–, and Arg–RNA Complexes

1. RNA fixes polar features of its ligands, often restraining them at the intersection of multiple directional bonds. Such restraint likely includes aromatic and heteroaromatic rings.
2. RNA can measure the distance between such polar features, possibly allowing substantial freedom in apolar bridging constituents by stacking them loosely.
3. RNA can also sterically limit the size, disposition, and/or shape of apolar groups close to the specifically bound polar elements cited in item 1.

Of course, other specificities may be added as more RNA-ligand structures are studied. This is particularly true because this analysis is based on only a few amino acid side chains and a few high-resolution co-structures.

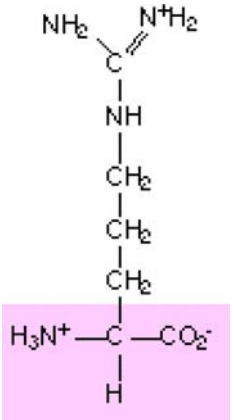
Part II: Selected Amino Acid Binding Sites

We now appraise the properties of selected RNA-bound amino acids. The lists below contain the most prevalent independent binding sites (each from a different parental molecule) recovered by selection for nine amino acid affinities, beginning with randomized RNA sequence libraries. In some cases the sites have been subjected to squeezed selection, reducing the size of the randomized RNA until affinity selection fails (Lozupone et al. 2003). These site(s), requiring the smallest space (minimal number of nucleotides), are generally in agreement with the most prevalent sites recovered when space is abundant (Legiewicz et al. 2005). Therefore, there exist multiple data implying that these sites (especially for Ile, Trp and His) are the simplest, most easily found amino acid binding sites composed of RNA. Squeezed selections, with differently sized randomized tracts also generate many new independent sequenced binding sites. These were not yet available during our last comprehensive review (Caporaso et al. 2005; Yarus et al. 2005). Thus, we can now assess an enlarged site library about seven times the previous size, making it likely that this review has greater resolution than any previous analysis.

In performing this review, we have several purposes. Even with no explicit mechanism in mind, binding of amino acids by RNAs is a possible source of primordial coding. The chemical properties of the interaction are therefore of interest, as indicated by the quotes at the head of this Chapter. Further, we can compare these newly known sites with the expected properties of a polar profile, to see if our generalization from recently known amino acid–RNA structures makes sense of selection data. If so, we may access a deeper understanding of amino acid

affinity, and even be able to anticipate some yet unseen interactions. In fact, amino acids that offer a more diverse polar profile appear to be bound to RNA more strongly and with more specificity.

Brief descriptions and sequence lists follow. The lists give independently isolated examples of selected RNAs, with initially randomized binding site nucleotides in upper case and the non-site nucleotides in lower case. Site nucleotides are defined by sequence conservation across independent isolations, and/or by chemical and enzymatic protections and sensitizations by amino acid ligands, and/or as well as by binding interference and facilitation after prior chemical modification. A (capitalized) site nucleotide is therefore protected by ligand from enzymatic or chemical attack, interferes with ligand binding if modified, or is independently conserved, or any combination of these properties. Non-site nucleotides in the same selected molecules (in lower case) have none of these properties, and serve as controls for our analysis (Knight and Landweber 1998). The site sequence files are the most complete presently available, and have been checked for accuracy, often against original data. Sequence curation has also been more rigorous; for example, cases where site triplets appeared to be forced by interaction with fixed flanking nucleotides have been eliminated. Sequence libraries below are available as text files on request.



Arginine

The side chain guanidinium ion has a resonant stacking pi electron system, a positive charge, and in addition a pattern of hydrogen bonding that matches well with the edges of nucleobases. Arginine thus presents a prototypical double ended polar profile, in which both α-carbon groups and the guanidinium terminus of its long side chain can interact with an RNA site.

The RNA sequences of arginine sites resist generalization, because they come from five different selections which employ varied methods and found varied binding sites in which no structure recurred. However, an NMR structure of one aptamer complex (Yang et al. 1996) is available, as is an arginine–TAR RNA complex (Pugilisi et al. 1993). The former complex shows a cage of nucleobases offering H-bonds to guanidinium, an aliphatic side chain stretched out across a purine nucleobase, then a simultaneous H-bond from α-amino to ribose, confirming the expected RNA focus on the two polar sites. The latter complex shows guanidinium stacked under one nucleotide and paired with the major groove face of G just below. This resembles the original arginine–RNA complex in the group I active center (Yarus and Majerfield 1992).

Different selected arginine sites range from $K_D = 0.33$ μM to 4 mM for the L-amino acid ($\Delta G° = -9$ to -3.3 kcal/mol), consistent with multiposition interactions with RNA sites. Further, the multiplicity of different binding site structures observed after selections is consistent with multiple opportunities for RNA–amino acid interaction; arginine is perhaps more flexibly bound than any other amino acid.

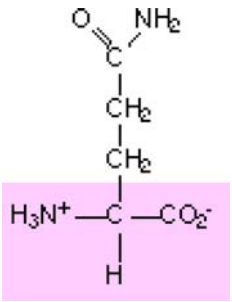
This flexibility permits the selection method to have a substantial effect the sites detected. A rigorous selection (Geiger et al. 1996) for side chain selectivity and slow dissociation yields likely double-ended sites with low K_D (=0.33 μM) and high enantioselection ($K_{L/D} = 12{,}000$). If selection is relaxed, likely single-ended sites with K_D of millimolar range are recovered (Connell et al. 1993; Tao and Frankel 1996). Single-ended properties are often obvious: some sites (Connell et al. 1993) make little distinction between L- and D-arginine, and bind guanidinium approximately as well as the complete amino acid. This can be of experimental importance because, for evolutionary purposes, the smallest, simplest, easiest to find sites are often sought (Lozupone et al. 2003). If an amino acid allows both double-ended and single-ended sites, it is the single-ended class of site which is probably emphasized by experimental selection for simplicity.

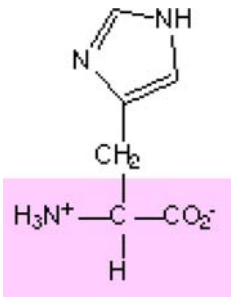
The site sequence list of arginine binding RNAs below contains ≈7 times as many independent sites and ≈7 times as many nucleotides as the previously analyzed site population (Yarus et al. 2005). A ">"marks the beginning of each sequence; after a line of identifying information, a line feed leads to the actual RNA sequence (lower case, non-site; caps, site nt).

Glutamine

This amino acid presents an RNA binding puzzle. Despite an obvious double-ended polar profile and a side chain amide that should readily and multiply H-bond to nucleobases, it is very difficult to isolate RNAs that bind the free amino acid. A partially randomized RNA had almost no selectable affinity for glutamine (less than four other amino acids (Famulok 1994)); and selections for L-glutamine

```
>Arg #1 reselection G. Connell, M. Illangasekare, M Yarus, Biochem 32:
5497-5502 (1993)
cgaaGCACcagcuaugugcaugGCC
>Arg-GMP G. Connell, M. Yarus, Science 264: 1137-1141 (1994)
cccgacagaucGGCAACGCcauguu GAGACACC
>Cit2Arg #6 - M. Famulok (seq via Rob Knight), JACS 116: 1698-1706
(1994)
augugacacCAGGUAGGUCGCcgcacgucgcuucaggacugugcgGAAGGAGCGGuagucaucaugcgguu
gc
> #7
ucgggaaguCAGGUAGGUCGCuggacagccuuucaggauaguccgGAAGGAGCGGc
> #8
uaaggauauCAGGUAGGUCGCuguacuucuaucaggacuguacgGAAGGAGCGGuauccuugcuacggauc
c
> #9
auggcaaacCAGGUAGGUCGCgggacauuucgucucGAAGGAGCGGuacgauugccaagggugua
> #14
auagguaacCAGGUAGGUCGCuggacuacacuuuagguuuguccgGAAGGAGCGGuuccuugcugcguaug
ug
> #16
ucggaaaacCAGGUAGGUCGCcggaugucacuucaggcauguucgGAAGGAGCGGuuuccg
> #17
ucgcgguggguucggugCAAGGAGCGGuuuaaaugcCAGGUAGGUCGgcaccgau
> #20
ucggcuguaagaauuggucgCAAGGCCGGuaguuacCAGGUAGGGCGgcaccgau
> #21
acccaauacCAGGUAGGACGCgggacuucuguucaggacugucccGAAGGUGCGGuagccuggguaaguau
ggc
> #22
ggggccaucCAGGUAGGUCGCuggauugauaugaguauuguccgGAAGGAGCGGaaaccucgcuaagcaug
gc
> #23
aacgcaaacCAGGUAGGUCGCuggacuugucuccuugacuguccgGAAGGAGCGGuauucuugcgauguuu
ggc
>Arg F2a"#2" J. Tao, A.D. Frankel, Biochemistry 35: 2229-2238 (1996)
uggugcgugcaggacGUCGAUcgaaUCCGc
>Arg F2b
augagucaccgccGUCGAGcuUCCGgcgug
>Arg F2c
ugacacacgauaGUUGAUacUCCUaucauc
>Arg F2d
ucguagguugaaCUGAUuUCCCuaaacgcu
>Arg F2e
gauagccgcguacaccaGUCGAUcgacaUC
>Arg F2f
acgcggcagcaGUUGACuaugUCCUgcacu
>Arg F2g
uuaGUCGAAguUCCUgagccuuuaagcccc
>Arg F2h
ugacccuaucuacuguagGUUGAUcuUCC
>Arg F2i
ugugcccuuucuaggGUUGACaagUCCCcg
>Arg F2j
acuccacagauggAUCGACauaUCUCcacu
>Arg F2k
acaacugaaucagguGUUGACcgggUCCAc
>Arg F2l
cauagacacGUUGACaaagUCCGuuccguc
>Arg F2m
ugccauccugugGUCGAGcauguuggcUC
>Arg F2n
aggacugaaGUCUGAGcggaugcUCCUucc
>Arg F2o
cucacaagaugaGUCCACgcaacgUCCUca
>Arg F2p
gugagccgggugcgaggAUUGACgcuUCUC
>Arg F2q
acaagagcagcuugGUAGACuugUCCCaua
>Arg F2r
uacaGUUGAAacUCCUguggccagaccgug
>Arg F2s
uagcucagugacacucaGUUUGACcuuugC
>Arg F2U
cUCCGguacgcGUUGAGaugcgugcac
> ag.06 (58 nt) Geiger, et al Nuc Acid Res 24: 1029-1036 (1996)
augauAAAccgAugcugggcgAuucuccugaaguaggggaagAguugucauguauggg
```





affinity in completely randomized RNAs often fail outright (Majerfeld et al. 2005). Nevertheless, low affinity sites have been obtained, in unpublished selections by C. Scerch and G.P. Tocchini-Valentini ($K_D \approx 2 \times 10^{-2}$ M; $\Delta G° = -2.3$ kcal/mol), and these are listed below. In the spirit of the polar profile, perhaps the polar sites of glutamine are at an awkward spacing for simultaneous interaction with an RNA site. In any case, this is in accord with evidence that glutamine entered the code by a second non-stereochemical route (Wong 1981; Yarus et al. 2005) not dependent on interaction with RNA.

```
>Glutamine 1  C. Scerch, G.P. Tocchini-Valentini,
unpublished
ucgauauuaaCACGGGUucuagcaaaagcucgugcugaccaucGGAUCAAGACG
uguugcccgacaaggguggcguggg
>Glutamine 2
acugGAUGUUCucugAUCGGGUaugcacucccuGGAgaaUCAAACG
cgugcaugcgauguuugagacccgguggugg
```

Histidine

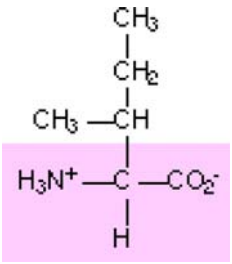
The predominant histidine RNA binding site (Majerfeld et al. 2005) is relatively simple, consisting of a hairpin with an adjacent internal loop. This binding site is highly selective for the side chain, for example, rejecting the other positively charged side chains on lysine and arginine. It is also sensitive to the protonation of the side chain imidazole, preferring the protonated form (the imidazole is unprotonated and uncharged as drawn above). Thus it may utilize stacking and hydrogen bonding to the charged imidazole side chain terminus, which would make it strongly double-ended in polar profile. These side chain specificities are accompanied by L-stereoselectivity of 100–900-fold in different isolates. These data suggest that the most readily formed histidine site is a double-ended structure like the lysine riboswitch (Fig. 4), which detects both side chain and α-carbon features. K_D for the most prevalent site is about 1.2×10^{-5} M ($\Delta G° = -6.8$ kcal/mol), again suggesting multiple points of contact with RNA.

The following site list has 4 times as many independent sequences and 4.5 times as many ribonucleotides as were available before (Yarus et al. 2005).

```
>His 945 I. Majerfeld, D. Puthenvedu, M. Yarus, J Mol Evol 61: 226-235
(2005)
aAAGUGGGuugAUGUaAGuAACAGgcgauaggcuuugcguuccaaauugcuaucuaacguuugcgcgcu
>His 240
AAGUGGGgugACGUaUGgAACAAcguuaguugcuuaggaacucucgguugguguucgug
>His 225
aaaagAAGCGGGgguaAUGUuuUGuAACAAcuuucuaaugagguaggagccucggauugcgugucgugu
>His 241
agcugagauccgaugggaugauaAAGUGGGugagGUGAaGGgAACAGaucguccuaucgugacaagug
>His 949
ucuaaUAGUGGGugacAUGGaAGgAACAGuagacagagagaggagagguucccaacgcgaauaaaggcuu
>His 206
augacAAGAGGGuaUAGUuAGgAACAGucauccugaugggagagungcuuacgugguccuaccau
>His 956
acggcaUAGAGGGaUUGUuAGuAACAGccacagauugagcagggaucgcaucguggauaggguuugucgc
>His 729
ggcauaauacaAAGUGGAuGAGUuAGgAACAGguauuuaugcaugguggaguucggguacacgucg
>His 1BD
gaacacaAAGUGGGAUGUaAGgAACAGgugaagaacgggagcugcaggguaaacgugaggcuguucaaug
>His 115D
ugAAGAGGGuaaAUGUgGGuAACACacugcggagcuugggauagcaucugaggcaagaagguguugccau
>His 103D
AAGUGGGguauuucaACGGaAGuAACAAcgauuagagaugggaaugucucucguuucgguugggauaua
>His 223
uggUAGUGUGagggaucugagcUAGUGGuAACAGagaaggaucuauaucagugcauucucaaguggggu
>His 14
caaucAAGUGGGuAAGUaAGcAACAGacugccgaagacggucggaucuugaaggcgcaaauggcgugg
>His 805
```

aAAGGGGGAUGUaAGgAACAGcccgauaaugcgaaggacacccuuugagaaagcuaggugauaggnuggg
>His 215
cuagucgggauuaguaauggagAAGUGGGuguaAUGGuGGuAACAAuucgucaaaaucuuacuuagcug
>His 253
cuaagguuaucaggaauuaaucAAGUGGGuugaAUGCgaGGgAACAGguugguccugguuaguca
>His 26D
aaucuugggcuucgaggcguccAAGUGGGuucaAUGGuAGuAACAGgacgucuucguacaaguguga cu
>His 164
gagagucgucucucagcaCAGUGGGgugAUGGuaCGgAACAGgcuggaaauggcguucucucc
>His 239
aaaacggaucuuccgaugaAAGUGGGguggagguuAUGAGaAACAUcgcggagugauucguucgcagacg
>His 31D
agguggccccucAAGUGGGaguAUGAGgAACAGggagcauccacccggauaggcaaccuguuuug
>His 235
gagguaggagcuguaugauaggaCAGUGGGaUUGGgAGaAACAGccuagaaugugacucuggcccug
>His 11
cguaauggucaAAGUGGGgugcggagucugcgugguagacccagcGGGUgUGuAACAGgaucugccgg
>His 15D
auaagucGAGUGGGaUUGUgAGgAACAGcuugucguaagggcuggauugggugcuauaucguaggua
>His 21D
ugucAAGUGGGuaauAUGUgGGaAACAGacggagugguggagcaaugguuccaacuauguauguaguucg
>His 243
gAAGUGGGaacAUGUgGGaAACAAccguaguagcgaucccgggauugcgaugaugguuagaagac
>His 9
gcaCAGUGGUgUUGGuUGgAACAGgcccuuuggagaacggauccuuaaguugcaucagaaguaugauc
>His 907
UAGUGGGuuUGGAuAGuAACACcgugaggcaaauggauuagccauuugagagcgauauggugcgcgc
>His 111D
ugAAGUGGGggUAGCaCGgAACAAgaguagccggauuacguuaaugaggguaacaaaccguaugcuguaa
>His 13BD
cuaccaUAGUGGGguUGGAaGGuAACAGgguugcugauagaugaugucgguuugggucucgguuacugcc
>His 321
uuaaccaggaguggaagugacgaAAGAGGGaaUUGUaAGgAACAGcgucacgauaacauacaggugaugu
>His 17
cagcAAGCGGGaaaAUGUuGGgAACAGcugcggaaggaaaucaugaggagcuuaugguucguugaaugg
>His 106D
aaCAGCGGGaguaAUGAaAGgAACAGuccgaggguguguauuagggauuauugacuggcgua
>His 3
cgAAGCGGGggggAUGGaAGgAACACgucgaugugacggauccagcacuugaucgugauugcguuu
>His 207
auggcaAAGCGGGGGGAaAGgAACAGgccauuguuaggacuuaggcagauggcuggucgguugu
>His 19
gcaAAGCGGGgGUGUuUGuAACAGgccuagauccggagccacgggauuaccccuacgauacgcuaccuua
>His 20
aucguugguugcuggagagguuagcuccuuacggAAGCGGGuAUGGuuAGaAACACcguaugg
>His 18
uaccguaugaGAGCGGGugUGGUcAGuAACAGccucgcgcuguuaguagcaugguuuccgcgaucgugac
>His 943
aacccagcggGAGCGGGAUGGugGGuAACACcgcggggugagauagaacuggacuuauguuaucuuguc
>His 817
aAAGAGGGuuagAUGAuAGuAACAUccgauaggcgaaacugacaucuaaugauggggcuaucccaugg
>His 370
uggugcugcacgaguuccaAAGAGGGgAGGUaAGuAACAGggagccaagugcgugcuucgcuagugug
>His 40BD
acaggAAGAGGGggaagAUGGaAGuAACACcuauucaauaguugcacaugguucccucgaucagccauc
>His 812
cgcugcAAGAGGGuAGGGuAGuAACAGcaguguuugcauaaaggngcgguauccuaaggagagcaggcu
>His 245
ccaAAGGGGGuggcAUGCaAGuAACAAggcgauagacaagauuucgacaauucgugaauggucuaaagcgg
>His 2BD
cguguugucacccguggaaauaagggcugggcgcaacagcaaAAGGGGGuauaAUGUuUGgAACAGugcg
>His 940
uggacuugagguacuccgugaauggaaucggAAGGGGGgGUGUgAGuAACAGccgaucugcucacggaug
>His 16
cggcgacaaucAAGGGGGguaAUGGaAGaAACAGguuguccgcaaccucgggcuuagaagugaagguc
>His 107D
ccuggugguggguaAAGUGGGuuauACGUgAGgAACAUagcguguuagccaggcggac
>His 944
AGUGGGuuuagAUGUgGGcAACACgaggugagcuuuuugcgaggcuucgguguaguuguauggg
>His 32D
ucacAAGGGGGuuaguuuguaacggugagcaggugauacgauacgCAGUaCAuAACAAcgauggagug
>His 242
ugggaaugagcaugccgugauggAAGCGGGuuUGGUuAGuAACUCcagaacggucaggaauugcac
>His 235B
acaCAGUGGGgucaggggcugcgcguacgguguguuacgucggcgCAGUGGaAACGUgucuugc
>His 38D
uugcguaugaggaugaauucuggggggguuaugAAGCGGGuuUUGAgAGuAACGCaugacccucucgc
>His 244
gcgacggggccuggggccaggcugaauugauccgucAAGUGGGgAUGGuaAGuAACGGgcgaauuagg
>His 231
ccuaaucCAGUGGGguuAUGUaAGaAACGGagagacacguuagaucuaaggguacucaggcccuuu

Isoleucine



Isoleucine's interaction with RNA is one of the most frequently explored of the amino acid interactions (Majerfeld and Yarus 1998). The predominant site is reproducibly isolated in the majority when RNA affinity for isoleucine is selected (Legiewicz et al. 2005; Lozupone et al. 2003). The binding region is an asymmetric internal loop that distinguished norleucine (same volume as isoleucine but a different shape) by 0.5 kcal/mol. The RNA site (Majerfeld and Yarus 1998) had $K_D = 0.5–1.2 \times 10^{-3}$ M ($\Delta G° \approx 4$ kcal/mol), and favored L- over D-Ile by ninefold (1.3 kcal/mol). These moderate affinities and selectivities, directed at both the α-carbon and side chain, seem consistent with isoleucine's likely single-ended polar profile.

The independent isoleucine binding RNAs in the library below are 30-fold more sites and nucleotides than have previously been available for analysis (Yarus et al. 2005).

>Isoleucine F1-1 / Il19 "AUU" I. Majerfeld and M. Yarus, RNA 4: 471-478 (1998)
auccgucgcGUUAUUGGGGUuccacuauaccagguucuuuugucugcagu
> Ile F1-3 "AUU"
gauaaucugcGUUAUUGGGGUguccuuaucacuuuugcucugcauugacg
>Ile F1-4 "AUU"
ugggcuguucuccgcGUUAUUGGGGUaccuccuccaacucugcugcauga
>Ile F1-5 "AUU"
ugggcuguucuccgcGUUAUUGGGGUaccuccuccaacucagcugucgug
>Ile F1-6 "AUU"
uacacGUUAUUGGGGUaucucuuacucgcuacaguagggcccucgggacc
>Ile F1-7 "AUU"
uagugacaucgcuuucUAUUGGGGaucguacuuacgacguuuuacgagag
>Ile 211 90 nt M.Legiewicz, C. Lozupone, R. Knight and M. Yarus, RNA 11: 1701-1709 (2005)
cguugugagggucgccguacgcgacgaucuagaggucaccacgcgauaaguaugggugcagccugUAUUGG
Ucggcagguucgu CACCcu
>Ile 250
cgcCUAUUGGGGCcugaugcgcguugggcaaguauacuugaccgacacgauaguucugucgaucaucccau
acagggcaugcaggUUACG
>Ile 259a
cgucacuguccgcagguacggucugauguuuaucagugcCUAUUGGGCcuaacggacgUUACGuauguggu
auuuacaugcuagugcgug
>Ile 831
ugaaugguuCUAUUGGGGaucgcguuagcgauUUACGgaccgcagauugggacuggaugugucaacaaggg
aaaaunggcgaaucccgua
>Ile 210b
uggugucuccacuuuguuaauuaggcgacuggcgcgcaucgUUACGGuucccgaaucauacucucggaggu
gggauccuauuggggcgca
>Ile 404b
guccugcguuuaauuucgacgugcgUUACGguugucagucacacagauuggcaagCUAUUGGGGugcgcgu
ugaguaacagauuuacacg
>Ile 226b
guccaucuggauaaacucccgcugcacguucuaucguuaagagccggcugCUAUUGGGGcaaaaugCUACGu
ggccagcuuauguuucggu
>Ile 817
ggaugggucgcaggacggcccguaaaauagcgccugagacugcggcuGACAuggcuugcggccuguugacgg
uUAUUGGUacgccuacaac
>Ile 235b
cagaacaaCUAUUGGGUgacuuacauauugcuaggacagcucagucaaaguggcgagcguccuuggcggac
aagucAUACGuuguuacag
>Ile 814
cuUUACGGaauacaugaugaaauuucCUAUUGGGGagcuuuuagagcauugguucagaccccgcaaacaca
uacgucaucgugcuuugcg
>Ile 237
uuaucaaccggcUUACGcguaugucaucacgagugugauuucgCUAUUGGGGgccgggaaacgugcauuca
gcgcugucauggagu
>Ile 244
cagcaccgggcguaggccucaacgccauuguCUAUUGGGCgcugacaguaagaaauuccgcGUACGgcaau
gagcguaugcaguggguu
>Ile 803
uugggcgcacguugcguuagaucauCUAUUGGGGcgacacaaacaacgauugcacauucguuugggucgUU
ACGgugaucacguguugc
>Ile 235a
gauugacgcgucauuccacuuugucggCUAUUGGGGccggggcucuaccgcgacguggaacucccccggUUA
CGcuguccgggaauacgc
>Ile 226a
cccaaaucacaugcCUAUUGGGGGucggguguuuggguuucuauguucggcaagacggagaucaaucuccu
accuguauccuggcuUACG
>Ile 835
acgugcggCUAUUGGGUuauuaggugggugauaguguauuAUACGuaagacggcuuaucguucgcgucaca
auucuaaugauacgucgca
>Ile 404a
uaguuuaaggguuggacuuaagaugaacCUAUUGGGGcuaauuggaauuuaaauuagUUACGguugaaucu
aucuuagcaacagcacugg

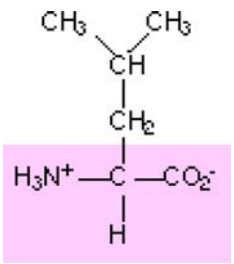
```
>Ile 406
ugccgauucaacuucuuagguccacaCUAUUGGGGgaaucuugucuugauucUUACGugugcaguauuauc
gcaacuccaaaauaucgg
>Ile 811
guaacucgaggUUACGauccgaauggacuuuucguguauCUAUUGGGGccucgagucgcggauuagguacu
agguauggggguaccuuagg
>Ile 245
auaaucacacagcuggUUACGaaacuguugcuaugaucugauggcaaugugcuuCUAUUGGGGcuagacua
gaugacuaucgcacuuccu
>Ile 407
accaggucaaaaccuuuucaacgcgguuacgcgcCUAUUGGGGGuccuugggguuucagcccagcacUUAC
Gacgcgccgcgugccuggc
>Ile 246
gucccgccugaacguagcccggguggaGUACGucacuuaauaguggCUAUUGGGCucuagcauuugcggag
gguuacgcucaggaucuca
>Ile 210a
ucgcguacuuacggaaaccaccggcCUAUUGGGGuccguuuaggaUUACGgcuggguugacgaaguguuuu
uggugugcucugcuaggc
>Ile 224
cggugccagccggggUUAUUGGGGaucacucguuaguaucggccuucguaccgggugcgaauuuucgauug
guUUACAccccaggcagac
>Ile 243
ggaaacggagcgaccugcgguucacugggucgcgcgucaauauuuuccgcuuuggugacucacguUACGgc
auugaugaagcCUAUUGGG
>Ile 827
gguuacaugagcgauuuacauaagcguaucaUUACAcugugaucggcguauagauuaggcacgaaccacga
auggcucaUUAUUGGGGcc
>Ile 451
UUACGcgcgcacgggaaauaugagaugaucaucggaggcgcgCUAUUGGGGGaagcaucuuagcauuuacg
u
>Ile 240
ucgcgcgCUAUUGGGGcccagagcuauucgguagacaagagaaguuuuaagguguaaacuggacgugucug
gcgcauuugcucugggUUAC
>Ile 258
acaccgguuGUACGuccugaggggCUAUUGGGCaccgguagcucugggguucuggauugguguucaacaaa
agauccaaugaaccguag
>Ile 259c
ugaugaagaacucgucaugguaaccuuugauaaugguucucuaucacaucagcaguUUACGggccagaaaa
augccCUAUUGGGGacgcg
>Ile 205
uuguguucuggcacccuugcacugcccacuaauuuuuuuuuuucaaaggcaaccgcgcggUUACGagcuaa
aucacuCUAUUGGGGucgc
>Ile 260
aauucaacacguCUAUUGGGGcgugacgcuuggcaucuagccaaagcauuaaugcgugcuauuguauuagg
gcugaguggucgcgUUACG
>Ile 429
gggaccCUAUUGGGGucuacgaguuguacaucgcuuaagaaauacucuucacgcguugaaguuugucaucc
aacuucuagaUUACGgguc
>Ile 816
ggccggagcaagugcuucacgccagagagcgcgaaaaucuaucgugccaggggCUAUUGGGUUguacuugc
accuccuggAACGaccuca
>Ile 202
aaugggcgaaaggcuaaccauuccacccgaggaccgcuuuacacggugcCUAUUGGGGugagcgcugaugu
gaaaggugaucaUUACGgc
>Ile 812
aagguugcUACGgcgggaugggaaaaagggcgggccuuuaaucuggaaaggcauccaauuuaccccuuguc
cagcgcCUAUUGGGGcuuc
>Ile 236 70 nt M.Legiewicz, et al. RNA 11: 1701-1709 (2005)
ucucggcgacaacgcguugacacgUUACAucuacucugacucuggggcguagaUUAUUGGGAuguguugu
>Ile 252
gacaauguguugccuauCACGUucaugacaacggaCUAUUGGGGuaggauaugaguucagcauuguuucc
>Ile 243
UACACuaguaccaaguaaugguguuagUUAUUGGGcgucucucuucauggucgggguuaaucggccc
>Ile 231A
gacaggacacacgcccagguaUACGguuaagugaguacuugaguucucgcaccUAUUGGGuacaauucgg
>Ile 404B
auugaguuucagcggcaCUACAGGucuuggguuauaauauuaauccaguCCUUAUUGGGGugcccucuga
>ILE 403B
AUCACUUACUUGGUUAUUGGGGGUUCCUUUAUAUAUCAUGGUUCUUACACCAAGUGGGUAGCCAUCGAGC
>Ile 417
uaguCUAUUGGGGaucgcgguauuaauuuuggguuauaccccgaUUCACGgcucuugguucguuccaguca
>Ile 225
cucguuaagugcCUAUUGGGGauaucugaucguguuuugagguggcaccggccggauauUUACGGcauga
>Ile 408
ccuaagugcaauuagaggaauauCUAUUGGGGccuaggucuaaucacaggacauuuaaggUUACGauguu
>Ile 414
auuggaguaaggggcucguggcCUAUUGGGGcaucuuaacgUUACGGcucgagcuuccucgaucugauuc
>Ile 206
uacccaggagcauuucugggagugccgcUACGccauuuuccgcugugucagugagauggCUAUUGGGGcg
>Ile 241
uCUAUUGGGGcuuaacagugaugucccucucaaaucaucuggugagCUACGacguuuuuccucuccgcag
>Ile 418
gucguguUUACGuauguuugacaccauuagaacaauaauuuggucacauucuaCUAUUGGGGacgcgacg
>Ile 214
aCUAUUGGGGcuucuuauugucaugaucauguuuauaaggagUUACGuaguacuugcucccucgucggac
>Ile 431
ggauugucgcugCUAUUGGGGugcuacauugccuucagaggaauguacaUUACGcaguuucgucgucuac
```

```
>Ile 432
ugguaggUUAUUGGGGcgguuguuuaaccucguaauUUACGauagaauguaacaccguuacaccuauucc
>Ile 440
guUUAUUGGGCcuucacaauguaauuuucuuauaugauggagaauuguuucaauuugugaacgGUACAau
>Ile 406
UUACAcgagcguguuucgugucgucuuugugaugacaaccgguugcugauacuuuagcucgUUAUUGGGa
>Ile 441
cgggcacucaCUAUUGGGGcgcaugccggcgguacuauauuugaguaaucggugcagCUACGuggaug
>Ile 422
cgcuccgaCUAUUGGGAcggaugauuugcuugauucuacuaguagcagcuccgUUACGuucggcuuuug
>Ile 436
uaaCUAUUGGGGagccgggugaauuuguuuaauucguagauaauuuuagcuuuucauuuugguuUUACGuu
>Ile 247
cgugagcugUUACAuaaguuguauugcgucugcguaaaagguagcuuaUUAUUGGGGcagcguaaggcgc
>Ile 231B
uaggccuuggcucuucgacuucuaaaguguaaUUACGucuguaaaggaCUAUUGGGGuuacuugccuggc
>Ile 202A
auaagucgCUAUUGGGGaucuaauuggauCUACGugacuagcuuaaaccacaaugcggauugccuccgca
>Ile 210
uuuacugugcaacgcaucgcaCUAUUGGGUcuauauuggAUACGugcgguuacaaucuccaucuucgcgu
>Ile 249
ucucuuugcucuuuucacgguaccacauuuauagcgUACGuuugauacuuaaaagcuaucgagCUAUUGG
>Ile 437B
UUAUUGGGGgucuggggcacagacgccauuuaguaguuuaaggauggcggcuuguugacccauucUUACA
>Ile 442
UUAUUGGGGcaaacacgugucgguccguuuacaguucaaggaacggguauagccgaaucgauuugUUACA
>Ile 426
aacuuuuaauuuugauaggguacgcuaCUAGUGGGGggcguguuuuuuagccUUACGuaguuuaccacag
>Ile 407
guugUUACGuuaggauuuuuguaacugggaauugcaauucacucucaucaaucccaaCUAUUGGGGcucc
>Ile 213
ugccuugaggugggaggaucauCACGccauggacuacaaauucugggCUAUUGGGAugauccugacacau
>Ile 416
gauacgccuggcguaaguacggUUAUUGGGGaagucacauuUUACAccguucccggagccuaauuucgga
>Ile 233
cgcgguGUACAgaguuccauccguuauuaagucaugucauggaaauauacgguggacucUUAUUGGGCcc
>Ile 424
cugcuuuaaggaaccuguguaacagugaucaacaCUAUUGGGCuuguuaaacgaGUACGuuguugcgaug
>Ile 427
gaCUAUUGGGCuacagucugugaacauuaguaaacggcaugagaaaugauuguacacagacgugGUACGu
>Ile 212
acagcgcagGUACGguucaacuaaagaucCUAUUGGGCcugcuaaggagccugcaagcccgauuguauga
>Ile 257
ugaaggauccguuucgcauugcaagaguugauaucuuaacacGUACGcguaaucauacuaacgCUAUUGG
>Ile 237
gcUAGUGGGGuuuuugaaguccagaugauaaauccaucaaggauccaaugucagcuguuagaaauCACGcc
>Ile 404A
ugguacaaauggaaccggcgaUUAUUGGGCuuucaguccggccgauugauuaGUACAucgucuccaagca
>Ile 202B
aaaacaugccagagguuagagucuaccGCACAuccggacagaggaUUAUUGGGCgguuuggucucggaag
>Ile 437A
cccaucuuucacCUAUUGGGGaauaugaguaaaguguaauaauauuCUACGugaaugauagguucc
>Ile 439
cuuucaugugCUAUUGGGGuuucggcaaccuuuaacucaaguaguugugucgagaUUACGcaaugggauc
>Ile 420
ugucuuuggCUAGUGGGCuugagucugguucgguuuaugaucuguaccaggugaaaGUACGccauauccg
>Ile 253
ugacgcacacccCUAUUGGACgguaugcuaggccuagcgcggguuaggagcgaggcccauugcucGUACGg
>Ile 208B
cgaaGUAUUGGGGgcgugcgcguaaucccccgcuaccgcaugagaucucuaggagcaaucgcccgaacggu
>Ile 433
AUACGuacgucaccaguucaugacggcguuagaucgccaaaguaugaacguuccgucgucaCUAUUGGGU
>Ile 123 50 nt M.Legiewicz, et al. RNA 11: 1701-1709 (2005)
ugGUACGaugggcucuugaugccgguagcuagaguucguCUAUUGGGCca
>Ile 127
ccguguccggugcacacGUACGcuuuaggacuuuuguucccggCUAUUGG
>Ile 305
ugggggggaucaaCUAUUGGGUcgaucauauuucgAUACGuuagguucccc
>Ile 314
ucaguucaUUAUUGGGCggaauguuucagcguuccGUACAugcacguagu
>Ile 134
ccaggcuggcguCUAUUGGGGuuguucaucagCUACGauugcggcuggc
>Ile 102
ucuggAUACGgugggguaaagguggguguuauuccgcCUAUUGGGUcuggc
>Ile 105
cgggaaaagagacCUAUUGGGcgggugcacggucguuuuuaauguccugc
>Ile 116
gccuugacaugaacgCUAUUGGGCcggugcauuuuuaagccgGUACGcgg
>Ile 120
ucuccauggcagccgagCUAUUGGGuauaagaguaCACGcucggcuuggg
>Ile 138
acacucuccgcgucagauucgccUUAUUGGAcauauuugUACAggcgacg
>Ile 310B
```

```
gguacccagaguGUACGcgcucaggucagagaucgCUAUUGGGCacucac
>Ile 106
cuucacacaauggCUAUUGGGGuuggacguuaugaagcaaccacgucauc
>Ile 341
uacgucccucgUUACGcaaaggCUAUUGGGGcgaaggggguguaagucuua
>Ile 320
cauaauuaaccacggcuagauugcgucuaguCUAUUGGGGuuagucuugc
>Ile 315
gUUACAccuacuacauuuauuuacgcuuaagauguauaggUUAUUGGGGc
>Ile 121
gauaugaaucuagcucacgcgaaaauugCUAUUGGGGgcuaguaugauccg
>Ile 124
auaggccuucacgcguuaucuguCUAUUGGGGugcuaagcaccacggcgu
>Ile 311
cgcgucaacCUAUUGGGGuuacaagauuaaCUACGgucauggacgcuucg
>Ile 336B
cgccuuaaaccucCUAUUGGGGaugaaaaagUUCACGguggaaauggugc
>Ile 136
acaUUACGugccugugagcCUAUUGGGGugucgcuuucugguugaccacg
>Ile 303
ggauccgcugAUACGcagucaugCUAUUGGGUcagcguugucuugugugc
>Ile 317
gaUUACAugacaggggggagcaauucccuacucugucaUUAUUGGGGucu
>Ile 302
gugacuugguguuggCUACGcgugucgCUAUUGGGGccggccuuuguuag
>Ile 140
uuggCACGuuugagcCUAUUGGGGcauauuaguguaugucacggcacagu
>Ile 310C
CUAUUGGAgaucagucacuaucaccggaacgguggauucuaaggucUACG
>Ile 109
ggaaucCUAUUGGGGuucgaauggcgcugucguucggaUUACGuguugcc
>Ile 335
cgggaacagguacCUAUUGGGGuaugauuggauuggucguacCACGguau
>Ile 144
UUACGgagucuuggcacuauccgucaguguacuucgagauucCUAUUGGG
>Ile 137
uucuuagcUACAucugccuauguacaaaagcggaUUAUUGGGGcaugauc
>Ile 307
gUUAUUGGGGucuugcuaaaaguguuauauuuucgagaucaagacUACAc
>Ile 327
ggccUUACGguguucuuucuauguguauuuuugagcagcCUAUUGGGGgg
>Ile 336A
UACGgaugauuguuugcucguuauauauucuggcaacacguuCUAUUGGG
>Ile 324
cauguCUAUUGGGGauacaggcacuacuauaguugaagguCUACGgcagg
>Ile 325
cggaacucuuauuuucuugcaggguCUAUUGGGcccacgguggcacgacc
>Ile 312A
AUACGuuacucugccguaauucuaaaggcaacagggaguuuCUAUUGGGU
>Ile 334B
cacgacguaggugcguguuuguaaggauaACGcggccgauguCUAUUGGG
>Ile 340
guGUACGuuucgccucaaggacucgguuguugguauagaCUAUUGGGCac
>Ile 308
auggugCUAUUGGGGuuuauccgcgcaaccuuuggauuuagaCUACGcgc
>Ile 337
ucgagCUAUUGGGGccugugaaaacgacuguaugcuuuuucacggCUACG
>Ile 112A
gggugagguaGUACGaauuucucaaugggguuCUAUUGGGCuaccagccau
>Ile 103A
CUAUUGGAccgucgacucggaccuagguccgcgaauuuuucgacgugUAC
>Ile 323
CUAUUGGGAuccacgagguucggagaaaaauccuaugccuguggaUUACG
>Ile Ile 14  26 nt M.Legiewicz, et al. RNA 11: 1701-1709 (2005)
UUACGagcgauguuagcuCUAGUGGG
>Ile 139
```

cUUACGagUACGaagcuCUAUUGGGG
>Ile 343
UUACGacuguaguuaguCUAUUGGGG
>Ile 7
uacuacuccguuuaggguUAUUGGGGu
>Ile 24
ucUUACAccgaagggUUAUUGGGGg
>Ile 129
cUUACAcccuacaugggUUAUUGGGG
>Ile 2
UUACGggcugugagagccCUAGUGGG
>Ile 345
UUACGgguguuuguuccCUAUUGGGG
>Ile 58
uuacaguguuccgugaacaCUAUUGGG
>Ile 124B
cUUACGuguuucgaacaCUAUUGGGG
>Ile 48
uucacacgUACGuguaaagCUAUUGG
>Ile 20
UUACGuguauugaacaCUAUUGGGGc
>Ile 317
uUUACGgucucaaggcCUAUUGGGGa
>Ile 347
UUACGuggcguaagccgCUAUUGGGG
>Ile 6
UUACGcggugaguugccgCUAUUGGG
>Ile 30
UUACGcggaugagagccgCUAUUGGG
>Ile 52
UUACGcggugcggagccgCUAUUGGG
>Ile 12
UUACGuccggaacaaggaCUAGUGGG
>Ile 304
UUACGuccgugcucggaCUAUUGGGG
>Ile 124A
cgUUACGuccguucuggaCUAUUGGG
>Ile 38
gcacgUACGguccuguggggcCUAGUGG
>Ile 40
cacgUACGuuccuuuguggggCUAUUGG
>Ile 11
GUACGuuguuagacagCUAGUGGGCc
>Ile 13
GUACGugcuuagugugCUAUUGGGCc
>Ile 336
GUACGugcaguccggcgCUAUUGGGC
>Ile 22
UUACGaucuuguuguggguCUAUUGGG
>Ile 53
UUACGcacuuuuguggugCUAUUGGG
>Ile 21
UUACGgaguuguauggacuCUAUUGG
>Ile 157
UUACGacuugaauugaguCUAUUGGG
>Ile 34
UUACGucguguugugcggCUAUUGGG
>Ile 306
UUACGgugcauuugcgcCUAUUGGGG
>Ile 301
UUACGgcguuauuucgcCUAUUGGGG
>Ile 17
cgUUACGcuacuugugCUAUUGGGGc
>Ile 350
gUUACGccgucuugagCUAUUGGGGc
>Ile 29
uccUUACGcuuggugCUAUUGGGGgg
>Ile 27
ucUUACGcauuaacugCUAUUGGGGg
>Ile 357
cgagcUUACGcguuugCUAUUGGGGg
>Ile 5
aaucgcgUUACGcgugugCUAUUGGGa
>Ile 126
uaucacgUUACGcauuagCUACUUGGG
>Ile 155
cacgUUACGccauuguggCUAUUGGG
>Ile 140
acaccgUUACGcuuuuggCUAUUGGG
>Ile 313
UUACGcuuuucacgaagCUAUUGGGG
>Ile 341
UUACGcucuuaauggagCUAUUGGGG
>Ile 19
UUACGcaucuugagggugCUAUUGGG
>Ile 31
cUUACGucauguuugugaCUAGUGGGG
>Ile 130
cUUACGuauauucuaCUAUUGGGGgc
>Ile 338
gUUACGaggucacgcuCUAUUGGGGc
>Ile 346
gUUACGcgaucaaacgCUAUUGGGGc
>Ile 348
gUUACGacauugacguCUAUUGGGGc
>Ile 354
gUUACGgcaugaguguCUAUUGGGGc
>Ile 358
gUUACGgucugaugauCUAUUGGGGc
>Ile 15
gCUAGUGGGcccugcugugGUACGcc
>Ile 46
ccgCUAUUGGGCucauggGUACGcg
>Ile 28
uccCUAUUGGGGccaagggUACGgg
>Ile 122
gcuCUAUUGGGccuggguugGUACGa
>Ile 123
uacCUAUUGGGccguuuugggCACGg
>Ile 352
guCUAUUGGGGucauuugaUUACGgc
>Ile 112B
CUAUUGGGGgauugcccauccUACGc
>Ile 311
CUAUUGGGGugucucaagacaUUACG
>Ile 23
guCACGguggacauguugcCUAGUGGGCc
>Ile 35
GUACGgugacgauugcCUAGUGGGCc
>Ile 332
UACGguucgaaUACGuucCUAGUGGG
>Ile 356
UACGgcacgcauaaguguCUAGUGGG

Leucine

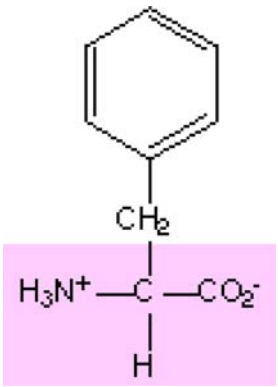


The leucine binding site was obtained alongside (I. Majerfeld, M. Illangasekare, and M. Yarus, unpublished) and by the same method of Sepharose-amino acid affinity elution as in the published phenylalanine selection (Illangasekare and Yarus 2002). The RNA family below was recovered 15 times (27% of the selected RNAs) and was the only isolate that responded to free L-leucine. The binding site is known from

S1 nuclease protection, Pb^{2+} protections, and DMS and CMCT base modification interference experiments. Leucine's aliphatic side chain should present a strongly single-ended profile for RNA binding. This is consistent with Leu 112's moderate $K_D = 1.1 \times 10^{-3}$ M ($\Delta G° = -4.1$ kcal/mol) along with an indistinguishable affinity for norleucine, but rejection of differently shaped isoleucine by more than 2 kcal/mol. Stereoselectivity against D-Leu is about fifty-fold (2.3 kcal/mol).

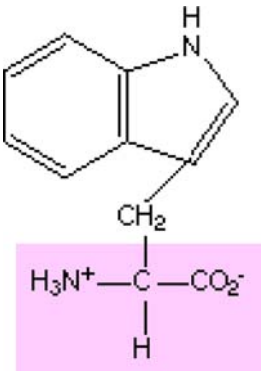
```
>Leu112   I. Majerfeld, M. Illangasekare
and M. Yarus, unpublished
ucucucAaccccUAgcgUAgUUUUGAcUGcGAGAGGCAAAcg
ccacggUAGAACCGAagGGUAGgagggauua
```

Phenylalanine




```
>Phenylalanine 529  M. Illangasekare, M. Yarus
J Mol Evol 54: 298-311(2002)
gcGcGAGAaacggucacuagaauaguggggccgUcAugcuaacg
ccucuuucgguguugGGGGAAUAUUggccaaucgagu
>Phenylalanine 523
auuggaucgguaguaUUuAGGGUGAGAcacuucaugccuuuguugca
ggcugggggugAAggcgcuacauggcgucUGAAA
```

The benzene in the Phe side chain can stack on nucleobases (Fig. 3, above), making phenylalanine an amino acid with a double-ended polar profile. Predominant selected Phe sites are three-helix RNA junctions with $K_D = 4.5 \times 10^{-5}$ M



($\Delta G° = -6.0$ kcal/mol), consistent with a multipoint binding profile. Some sites express side chain selectivity, distinguishing phenylalanine, tyrosine, and tryptophan side chains, and are highly stereoselective as well (Illangasekare and Yarus 2002), confirming the potentially double-ended focus of an RNA binding site.

Tryptophan

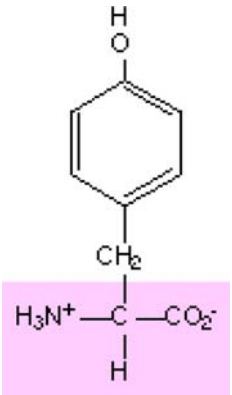
The large heteroaromatic side chain should stack and form polar bonds. Thus tryptophan is an example of a double-ended amino acid that can be bound to RNA via both the α-carbon groups and side chain. This is consistent with frequent RNA sites that bind several aromatic amino acids, but not other types of side chains (Majerfeld and Yarus 2005; Zinnen and Yarus 1995), consistent with a focus on the aromatic grouping. These sites can relatively have low K_D (e.g., 12 μM; $\Delta G° = -6.8$ kcal/mol) for simple RNA binding structures consisting of a small symmetrical internal loop (Majerfeld and Yarus 2005). The predominant and simplest tryptophan site selected is particularly sensitive to α-carbon substitutions, and also requires the heteroaromatic indole side chain grouping, as would be predicted for a two-ended polar profile (Majerfeld and Yarus 2005). Further, the $\Delta G°$ suggests the formation of several substantial secondary bonds, consistent with this discussion.

Tryptophan binding RNAs in this library increase the number of independently isolated sites by almost 5-fold, and the length of sequences available by almost 4-fold, compared to that previously analyzed (Yarus et al. 2005).

>Trp 70-93 I. Majerfeld and M. Yarus, Nuc Acids Res 33: 5482-5493 (2005)
AAGACCGucgcgcaagguuuugugcgguaCGCUACUUcgaggcugugggcaucucgggugugccauga
>Trp 70-305
gacUGCUACCAuacgcggggauugcguagaGGGACCAgacgaugcgcugaggaccugauuuaucaaagcc
>Trp 70-317
gacCGCUACCAcagggcuguGGGACCGgcgguagagauguagugacuuccccauacuggauaacuagcgc
>Trp 70-730
GGACCGucacgggaugaauaguauggcuguggaaCGCCACCCagcaugggcgggauccugaucgau
>Trp 70-727
cuggacgacggggaCGCCACUGgacuagguaagccAGGACCGuacgucgggagccgucagaaua
>Trp 70-151
guaCGCUAAGGuagguguacgCUGACCGuacgauaaagcuuagacguuccuagaaucaggacgcgg
>Trp 70-91
guuaauaagaccucggaggaguuagggucaauucggcauaggcugcugAGGACCGuaaccaguCGCUAC
>Trp 70-92
GACCGggacuggguuuuccacaguuggcCGCUACCGacuagaagcgaaauucccgaacgcuguaggcu
>Trp 70-372
guuCGCCACUCacagcgcugagcgugGGGACCGacgauagggggguuuuggggugagauaugaagcuuggggc
>Trp 70-148
ggcCGCUACUCgguuggggguuaucgAGGACCGggauagagcuaaggaugggguucguuguc
>Trp 70-156
gcacgagGGGACCGggugcuggaugagcaugacCGCCACUCucgaccagaucgggcuaugaau
>Trp 70-398
gguuacuuuguggggacgAGGACCGugaccuuacguaagggaguCGCCACUCguucgaguuauuguacgc
>Trp 70-381
uucuaaguggggcggAGGACCGucauucagagcagauggaugguaCGCUACUCuucucacggggaacugc
>Trp 70-553
gaaCGCUACCCugugguuggagcacgagGGGACCGucgauagaguaguaguguauccuagugcgagaccgu
>Trp 70-40D
cgauaggcacacuucugGGGACCGuugaaGCCCAUGUgguucccguguuugcuugcggagcga
>Trp 70-354B
gnaggaauggcuugcgGGGACCGuguuuaaguCGCUACCUgcuugucgguuucccguaunccuu
>Trp 70-711
gggaaggagcguccuuauguaccggcgcaaugguugCGGACCGcggcgaggacgguauCCCCAUCAcg
>Trp 70-360
gGAGACCGggguauuuugaaagggaugucCGCUAUUAacgauagaagcucggagacaaugacggggaugga
>Trp 70-117D
AACCGngugcgacuuaaugcaugaaCGCUACGAgggaguaaugaguuaguggaacuggguugcuuguguc
>Trp 60-203
uagcccaugucCGCUACGcaaguaaguugguaagauuugaauggacUGGACCGgguggggg
>Trp 60-204
gaugcuuaaucuggcggguauguuCGCUACUcacaaggugAGGACCGugacccggugg
>Trp 60-218B
uaauguugccguaggcgcaGCGACCGuauucaguCGCCAGUuguggcuguggcaaggugu
>Trp 60-207
uacagcgaucguggGGGACCGuggauuucauaucauguuCGCUACCcggcgcgcucg
>Trp 60-236
gggcgucCGCUACCuaacgcucgaguugagagucaauagagguuguuguGGGACCGggcc
>Trp 60-223
ggucCGCUACCacugcucaaguGGGACCGgcaacacgacuaguguuucgaccauuaacac
>Trp 60-209B
guguauagaagguggacCGCUACCucacaauuuugggGGGACCGgcaccucgauaacgau
>Trp 60-231
ugcgucacgauggcaaccgGGCUACCuguaGUGACCGcggccaucaaagugacggcgggg
>Trp 60-229
augcagcgacgggaggucgacCGCUACCugagauaaGAGACCGgggccuuccguugccgg
>Trp 60-202
ugcacaauguucaaguuCGCCACCucuuuagacauggGGGACCGuuugugcauuggugau
>Trp 60-285
uggucCGCUACCacguguGGGACCGgcaaucugggccguugcuaugggguucugagguugc
>Trp 40-102
cgguguauggGGGACCGuauccaguCGCCACCccuuaugc
>Trp 40-127
gcugcuuuugaugucCGCCACGguuuugcCGGACCGgguc
>Trp 40-112
uuaggucCGCUACCguuugucguggacauuaGGGACCGgc
>Trp 40-122
ucguaggguCGCUACCacaaagguGGGACCGgcuuacug
>Trp 40-164
cgguguauggGGGACCGuauccaguCGCCACCccucgugc
>Trp 40-156
cgguguauggGGGACCGuacccaguCGCCACCccunaggg
>Trp 40-120

```
uAAGACCGgcgccauuuacucguauauggcgggcCGCUAC
>Trp 40-107
uuUGGACCGggagaguuuugugucacucuugucCGCUACA
>Trp 40-111
ucacuuuaggucUGCUACAuguguuaUGGACCAgcugaag
>Trp 40-145
uggcaauguuuCGCUACUgagauuagucAGGACCGugugc
>Trp 40-137
gcaagguaCGCCACCuggcguuguccaGGGACCGucuugg
>Trp 40-143
ugguaugucCGCUACUguugcguguaacAGGACCGgguuc
>Trp 40-129
agaugucCGCUACAggucgauugaugaacUGGACCGgguc
>Trp 40-141
cuuaugacCGCUACCuuauacgucugcguaaGGGACCGgg
>Trp 40-123
uagugGGGACCGacacuuuaagcauuugggCGCUACCuac
>Trp 40-105
uagguuCGCUACCuggaacggauuaacuccgGGGACCGuc
>Trp 40-166
agcuuguaCGCUACUaccauagauguuauugguAAGACCG
>Trp 20-636
GACCCucacguugaaGGCUAAG
>Trp 20-625
gAGGACCGguacggcCGCCA
>Trp 20-642B
gAGGACCGguucggcCGCUA
>Trp 20-616C
gAGGACCGuauaaggCGCUA
>Trp 20-637
gAGGACCGuaaaguuCGCUA
>Trp 20-613
gAGGACCGguacggcCGCUA
>Trp 20-607
GACCCucucguugaaGGCUAAG
>Trp 70-585
uugggcgacgggaucuacgaagucaggucgugguacaacgAAACccucUaguugaaggcuAacggaaug
>Trp 70-358
gaagCagUagAgggAcuuccgcuccgaauaggggcgcgaaggauggaguaaggauucaguacg
```

Tyrosine

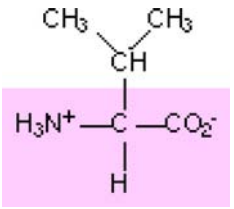


The phenolic side chain of tyrosine makes it double-ended in polar profile, potentially offering an RNA ring interactions including stacking, and H-bonding to its side chain hydroxyl as well. In fact, it proved easy to convert dopamine-binding RNAs to L-tyrosine sites (Mannironi et al. 2000). The RNA sites are hairpins adjacent to a helix junction, and maximally bind L-Tyr with $K_D = 23$ μM; $\Delta G° = -6.4$ kcal/mol). These sites prove to be L-stereoselective by 11-fold, and to require the side chain hydroxyl for best affinity, confirming a moderate double-ended specificity profile (Mannironi et al. 2000).

Valine

The prevalent valine site in RNA is an internal loop, 4 over 10 nucleotides. Its derivation did not permit deduction of RNA site nucleotides, so we have not used it below for coding triplet calculations. Interest in its original detection (Majerfeld and Yarus 1994) was instead focused on data suggesting that it preferred L-valine and could distinguish amino acid side chains that differed from valine by one methylene group by up to 1.6 kcal/mol. This raised the unexpected possibility that RNA sites could distinguish aliphatic hydrophobic structures by interacting with them productively. However, the polar profile puts this observation in a new light.

```
>Tyrosine 1  C. Mannironi, et al RNA 6:520-527 (2000)
ggcAGucaacucgugcgaucgugaaaAcGGGGcaAGAuGGccuuAcaGCGGUCAAUACGGGGGuCAuCAGA
UAGGGAGGCCUcCUggu
>Tyrosine 2
gagcgacugagguucgccgcggAUUAUguuuugcgguuagAUcaggcaaCGGGUAAUACcGGGuCAGUCAG
AUAGGGAGGauCUaCUgcc
>Tyrosine 3
aagggcaguccCCCCUucgcuggggggguGGuUGUAGggcuaaaacaaaccaCGgGUGAUACGgGGGcCAu
ACCCUAGGaAGGccCUgCUccc
```



We would now reinterpret these observations in terms of the polar profile of L-valine, which would predict binding of its α-carbon polar groups, and detection of the shape and/or volume of the adjacent aliphatic side chain. This new interpretation is supported by data emphasized in the original discussion (Table 2, (Majerfeld and Yarus 1994)), showing that the 1.6 kcal/mol distinction only occurred when a methylene group was moved or removed on L-valine itself. The distinction made between smaller side chains differing by a methylene was considerably smaller. This suggests that the RNA site wraps around bound valine, and the cognate valine side chain allows the formation of an optimally stable surrounding RNA structure. Thus perturbations (±methylene) from or to valine's shape alter stability of the complex, but the same perturbation far from the correct side chain shape and volume has less effect. In this way, the valine site acts as predicted from polar profiling. The relatively low net free energy of interaction ($K_D = 1.2 \times 10^{-2}$ M; $\Delta G° = -2.7$ kcal/mol) is consistent with a single-ended polar profile emphasizing α-carbon substituents.

Conclusions for Selected Amino Acid Sites

Viewed now from the vantage provided by the above 337 independently derived sites containing 18,551 total nt, with 4,945 nt of these within sequences essential to amino acid binding activity (sites), selection for RNA affinity for these nine amino acids has a roughly predictable outcome. Amino acids that offer and are bound via a double-ended polar profile, so that both ends of a bound ligand are caged, can be bound to RNA with $K_D \approx 10^{-6}$–10^{-4} M, $\Delta G° \approx -8$ to -5.5 kcal/mol.

However, outcomes will vary with the rigor of the selection, even for a particular amino acid. In particular, double-ended polar profiles can decline to single-endedness. This is well illustrated in selections for arginine affinity. A potentially double-ended profile can yield only a single-ended site in a selection for low affinity, for example, when smallest RNA sites are sought. These single-ended sites still distinguish side chains satisfactorily, and so are still potentially relevant to coding. In particular, they might still contain unique cognate triplets, which must differ between amino acids and therefore would likely participate selectively in side chain specific, rather than α-carbon, interactions.

Aliphatic amino acids that offer only single-ended polar profiles bind with about half the free energy of double-ended ones, $K_D \approx 10^{-3}$–10^{-2} M, $\Delta G° \approx -4$ to -2.8 kcal/mol. A crucial point for coding is that, under roughly physiological conditions, both classes of sites show sufficient affinity to bind amino acids from very dilute solutions, and in doing so, can exert considerable stereoselectivity under 'biochemical' conditions. Selections on the uniquely recalcitrant L-glutamine offer a known exception to the above classification.

A translation system made of RNA also must show chemical selectivity (or there will be no coding). In the reviewed RNA sites, there usually is side chain selectivity of several orders of magnitude, though off-target affinities can be so small for mismatched amino acids in single-ended sites that they are difficult to measure. Enantioselection is onefold (no selection; 0 kcal/mol) to several tens-fold (ca. 2 kcal/mol) in single-ended sites, and is 10- to thousands-fold (ca. 1–5 kcal/mol) in double-ended ones. Like enantioselection, all forms of selectivity against congeners tend to be greater with potentially double-ended, rather than single-ended, polar profiles. This is probably because single-ended sites, particularly for aliphatic hydrophobic side chains that must emphasize α-carbon polar groups, discriminate side chains *indirectly* by yoking an adaptive RNA site confirmation to an embedded side chain shape and size. For some purposes it may be important that binding which distinguishes only slightly between different aromatic (Phe, Tyr, Trp), cationic (Arg, His, Lys), or similar aliphatic side chains (Val, Ile, Leu) is also known. That is, known RNAs could also support ambiguous translation that embraces chemically similar amino acids (Fitch and Upper 1987).

Therefore, RNA sites can easily bind amino acids or carboxyl-activated amino acids (see "Part IV: A Model"), making sufficient distinctions among them to support coded peptide synthesis, in which a pre-coded stereoisomer and side chain selectivity would potentially be emphasized at each encoded position. Though it could not be obvious beforehand even to prescient observers like Carl Woese or Leslie Orgel, amino acids interacting with RNA, acting alone, can support specific, potentially code-forming interactions.

The above comments, perhaps surprisingly, greatly underestimate the *best* RNA sites. Instead, our conclusions are apt for the most easily isolated; that is, the *simplest* possible RNA sites. For example, amino acid binding by riboswitches employs larger RNA structures which can and do bind more tightly than our selected sites, presumably because riboswitch binding must generate free energy required to drive accompanying regulatory reactions. This is as anticipated—larger RNA sites bind GTP better (Carothers et al. 2004) because larger sites are better pre-structured for nucleotide binding (Carothers et al. 2006), rather than because larger sites make more interactions. Though amino acids are somewhat different double-ended ligands with loose linkage, unlike nucleotides, a similar progression might be anticipated, and has been partially characterized for arginine (Geiger et al. 1996).

We now proceed to analysis of the site-selected sequences listed above, which requires specific quantitative comparisons via computation.

Part III: Evaluation of the Distribution of Coding Triplets Around RNA Sites

Shortly after specific free amino acid binding sites on RNA were discovered (Yarus 1988) in the Group I active center, it became clear that bound arginine was associated with a site containing conserved arginine codons (Yarus and Christian 1989). Such interaction potentially provided a way to bring together RNA triplets and their cognate amino acids in a stereochemical origin for the genetic code. Therefore, we have embarked on an extensive search to see if varied RNA binding sites could be found via in vitro selection, and if they would embody the same triplet side chain relation. This current survey addresses 24 cognate codons and 24 anticodons potentially associated with 8 amino acids bound in 337 independently isolated binding sites. It is the latest and broadest published test of the concentration of cognate triplets in binding sites. Accordingly, we stress new results, leaving older discussion to reference (Yarus et al. 2005).

The relevant prediction is that coding triplets will be unexpectedly frequent in cognate RNA–amino acid binding sites. As a null hypothesis we assume that cognate coding triplets are equally frequent everywhere, inside and outside RNA binding sites. We test this hypothesis by calculating G (with the Williams correction) for the experimental sequences (G is a log likelihood function of the ratio of observed to predicted abundance, assuming equal abundance inside and outside sites) that is distributed as chi-squared (Sokal and Rohlf 1995). Because our prediction is directional, we test whether there is a higher than expected proportion of triplets within binding sites. The expected value for G is zero for a random distribution, and as triplets become unexpectedly concentrated, G will increase. Therefore large G and small P mark a significant triplet concentration. Each individual site is compared to its own flanking controls.

Triplets and Sites

The most easily found, simplest RNA binding sites for eight amino acids have an exceptionally improbable property related to the genetic code. Nucleotides essential to amino acid binding function include an unlikely number of cognate coding triplets (Table 1). Similar results have now been obtained using different statistical methods (compare (Yarus et al. 2005)), and so this conclusion is robust to changes in the analysis. By current reckoning, the probability that cognate coding triplets are *evenly* distributed with respect to their binding sites is $P_{\text{Codons}} = 5.3 \times 10^{-45}$ and $P_{\text{Anticodons}} = 2.1 \times 10^{-46}$. Thus both cognate codons and anticodons are very decisively non-random with respect to amino acid binding sites. These aptamer methods produce negatives as well as positives; note that by the current calculations, neither leucine nor glutamine sites significantly contain either kind of triplet. Thus the above P_{Codons} and $P_{\text{Anticodons}}$, while very impressive with respect to normal statistical testing (where $P \approx 5 \times 10^{-2}$ to 10^{-2} are often taken as significant), are even more so because they give negative results their full weight. In the present analysis, we have counted each independent isolation of a site as a separate event, whereas in previous analyses we generally counted only the initial site. However, tests of independent isolates with the same conserved sequences confirm that they are functionally the same; they bind with similar affinities and specificities (Illangasekare and Yarus 2002; Majerfeld et al. 2005). Thus counting new independent isolations seems justifiable on first principles, and both methods give concordant results (below).

Reproducibility

These results resemble those from previous statistical methods (Caporaso et al. 2005; Yarus et al. 2005), obtained

Table 1 Probabilities that cognate coding triplets are unconcentrated in sites

AA Sites/tot nt/site nt	Codon	P_{Codon}	Corr P_{Codon}	Comple anticodon	$P_{Anticodon}$	Corr $P_{Anticodon}$
Arg 34/1443/461	CGU	0.92	1	ACG	0.85	1
	CGC	0.0014	0.017	GCG	0.0018	0.022
	CGA	0.59	1	UCG	2.8×10^{-6}	$\mathbf{3.4 \times 10^{-5}}$
	CGG	3.4×10^{-4}	$\mathbf{4.0 \times 10^{-3}}$	CCG	0.78	1
	AGA	0.72	1	UCU	0.71	1
	AGG	1.2×10^{-20}	$\mathbf{1.5 \times 10^{-19}}$	CCU	0.71	1
Gln 2/156/42	CAA	0.042	0.16	UUG	0.97	1
	CAG	–	–	CUG	0.95	1
His 54/3644/969	CAU	0.87	1	AUG	0.010	0.039
	CAC	0.12	0.40	GUG	1.6×10^{-8}	$\mathbf{6.4 \times 10^{-8}}$
Ile 185/9915/2508	AUU	8.0×10^{-110}	$\mathbf{4.8 \times 10^{-109}}$	AAU	1	1
	AUC	1	1	GAU	1	1
	AUA	1	1	UAU	3.2×10^{-131}	$\mathbf{1.9 \times 10^{-130}}$
Leu 1/73/37	UUA	0.98	1	UAA	–	–
	UUG	0.029	0.30	CAA	0.71	1
	CUU	–	–	AAG	0.95	1
	CUC	0.99	1	GAG	0.25	0.97
	CUA	0.30	0.99	UAG	0.006	0.07
	CUG	0.30	0.99	CAG	–	–
Phe 2/160/35	UUU	0.98	1	AAA	0.012	0.047
	UUC	0.98	1	GAA	5.5×10^{-5}	$\mathbf{2.2 \times 10^{-4}}$
Trp 56/2889/763	UGG	1	1	CCA	2.7×10^{-13}	$\mathbf{5.5 \times 10^{-13}}$
Tyr 3/271/130	UAU	0.026	0.10	AUA	6.0×10^{-6}	$\mathbf{2.4 \times 10^{-5}}$
	UAC	0.0041	0.016	GUA	0.0020	$\mathbf{8.0 \times 10^{-3}}$
Sum 337/18551/4945	$P_{Codon} = 5.3 \times 10^{-45}$			$P_{Anticodon} = 2.1 \times 10^{-46}$		

P is the probability that cognate coding triplets are not elevated inside sites, compared to non-site nucleotides in the same molecule, using a two-tailed *G*-test with the Williams correction. Corr *P* is this *G*-test probability with correction for multiple sampling across triplets (calculated using Corr $P = 1 - (1-P(\text{single}))^N$ for *N* triplets for each amino acid). Corr $P > P$ because it is always more likely to find a relation in multiple trials. A dash indicates that a triplet did not occur in the experimental sample. As before, we take Corr $P \leq 0.01$ to be significant deviation from a uniform triplet distribution across a full set of triplets, and have emphasized these cases with bold face. P_{Codon} and $P_{Anticodon}$ in the Sum line reflect combined data for all codons or anticodons using Fisher's method (Sokal and Rohlf 1995) for independent experiments, applied to *G*-test probabilities for complete sets of triplets sought together in cognate sites

using the then-current sevenfold smaller sequence sample. In particular His, Ile, Phe, Trp, and Tyr previously had significant anticodon concentrations and Ile and Arg previously were cited for exceptional codon concentration in binding sites. Gln, again as before, has no significant tendency to elevated triplet frequency. There are three new results: first, that leucine now also has no significant triplet concentration. Because the single-leucine-binding RNA is unchanged in previous and present tests, these changes are attributable to more conservative site definition and statistical testing in the present work. Second, Tyr codons narrowly miss significance. Lastly, in a much larger set of binding site sequences, an arginine anticodon (as well as codons) is now seen unusually frequently in binding sites.

Two Kinds of Triplet Concentration

The changes cited in paragraph 2 simplify the overall result; particularly the observation that a larger sample of sites suggests that arginine sites contain both anticodons and codons. There now appears to be no amino acid associated with its codons only; either sites contain anticodons only (His, Phe, Trp, Tyr), or they contain both kinds of triplet (Arg, Ile).

Sparseness of Triplet Usage

With this larger sample of sites, we attempt to resolve the contributions of individual codons and anticodons (Table 1). Our eight amino acids potentially employ 24

codons and 24 complementary anticodons of the 64 potentially once devoted to amino acids. Of the possible individual triplets, only 3 of 24 codons and 7 of 24 anticodons are significantly found within amino acid binding sites. Thus use of triplets is sparse, as one might perhaps expect—only certain triplet sequences (21% of total) occur disproportionately within functional binding sites. We can therefore name arginine CGG-AGG/UCG; histidine -/GUG, isoleucine AUU/UAU, phenylalanine -/GAA, tryptophan -/CCA, and tyrosine -/AUA-GUA as best candidate codon/anticodons for participation in a stereochemical era of genetic code assignments based on amino acid binding.

Negative Controls

For the same reason, at this higher level of resolution, we repeat a conclusion drawn in paragraph #1 above: a majority of these experiments (e.g., 79% of specific triplets) have negative outcomes. These can be taken as negative controls, suggesting that these procedures are not strongly biased to find triplets in some profoundly cryptic way.

Two Kinds of Sparseness

Concentration on certain triplets can be explained in two ways: firstly, for a hydrophobic amino acid like isoleucine, with a single-ended polar profile, sites are of a small number of kinds, and selections are invariably dominated by variations of a single site, which can comprise 90% of selected sequences in vitro (Legiewicz et al. 2005). Here sparse and improbable triplet usage reflects the underlying recurrence of a particular site sequence, because one particular site is more probable (more frequently isolated, simpler) than others. A sparse result may have a different status for easily interacting arginine, which possesses many triplets embedded in many kinds of sites; so far, arginine sites do not recur in independent selections. Amongst this variety, the fact that 4 of 6 arginine codons and 5 of 6 arginine anticodons are not significantly concentrated in sites (Table 1) supports the idea that selection experiments discriminate between triplets that can easily participate intimately in site structures and those that do so with difficulty. Of course, the same idea may explain isoleucine site behavior, but the conclusion is less clear there.

Missing Triplets

The observed sparseness in stereochemical triplet usage raises a further question: how did the missing 79% of triplets (Table 1) enter the code? In looking for the answer, we accept a stereochemical era, and peer through it to the mechanisms behind its events. We still hold the opinion we have expressed before (Yarus et al. 2005)—many triplets probably were added to a stereochemical core by coevolution with metabolism (Di Giulio 1999; Wong 1975), and by adaptative selection to reduce the impact of errors (Freeland and Hurst 1998). Both routes require a core of codons (perhaps from stereochemistry), and both have independent support: the intermediate misacylated aminoacylated aa-tRNAs required by coevolution have been detected (Sheppard et al. 2008), and the genetic code shows persuasive evidence of optimization (Freeland et al. 2003). We have explicitly shown that a stereochemical core is quantitatively consistent with later code optimization (Caporaso et al. 2005) and that triplet appearance and adaptation are not causally interrelated.

We accept the suggestion (Koonin and Novozhilov 2008) that random assignments in the manner of Crick's "frozen accident" (Crick 1968) are also consistent with concurrent stereochemistry, co-evolution, and adaptation; all four together may have shaped the ultimate 'universal' genetic code. Our best current summary of the implications of these data relies on multiply recurring trends that are unlikely to be radically revised by further experiments—a majority ($\approx$6/8, Table 1) of amino acids appear to have participated in a stereochemical era of coding assignment based on RNA-binding sites (Table 1; (Yarus et al. 2005)), but a minority ($\approx$10/48, Table 1) of codons and anticodons for participating amino acids were directly assigned via such stereochemical associations.

Stereochemistry and Complexity

Finally, it is sometimes thought to be surprising that amino acids like arginine and tryptophan, which have complex biosyntheses, are found to belong to the stereochemical group. Confirmation of these prior assignments in a greatly expanded analysis of these particular amino acids (Table 1) resurrects this question. However, we do not think these findings raise a new or difficult point. Firstly, replication of RNAs accurately so as to preserve ribonucleotide sequences is among the logical necessities for the evolution of coding and translation. Thus highly organized nucleotide synthesis pathways and energy metabolism must have existed in the environment that saw the development of translation; it seems to add little new complexity to impute a concurrent pathway for synthesis of arginine or tryptophan. Secondly, when little information is available it seems to us particularly important to follow the data (Table 1), rather than preconceptions for which experimental evidence is absent.

The Overall Hypothesis

The overall probability that cognate codons are not concentrated in amino acid binding sites now can be estimated

at 5.3×10^{-45}; the probability that cognate anticodons are distributed independently of amino acid sites is yet smaller (c.f. Table 1), 2.1×10^{-46}. Both kinds of unbiased triplet distribution are vanishingly improbable, by R.A. Fisher's method for combining independent experiments. Because what is combined in these calculations are probabilities for different sets of triplets in different molecular site populations, the underlying numbers are definitively independent, as required for the conclusion. Thus, there is no doubt that cognate coding triplets are disproportionately present in the simplest RNA-binding sites for amino acids.

Specific Criticism

Ellington (Ellington et al. 2000) has criticized a prior analysis. However, (1) this criticism applied only to work on an initial set of arginine sites, which has been greatly expanded. Even on that narrow basis it is not self-evidently correct (Knight and Landweber 2000). (2) Statistical criticism relied on tests that did not evaluate the essential idea that predictable triplet nucleotide sequences should appear in binding sites. (3) Some criticism was based on results with arginine peptides—as seen in parts I and II, peptides necessarily present a unique single-ended polar profile to RNA, and thus do not appear the same to RNA as free arginine.

(Koonin and Novozhilov 2008) seemed also to rely on arginine results alone, and were apparently under the impression that RNA–amino acid interaction is too weak to be evolutionarily functional. This completely mistakes the data, summarized in Part II of this review. It appears to us that thus far, no critic has tried to grapple with the breadth, robustness, or variety of amino acids in which code triplets and cognate RNA sites have now been found to be interlinked.

Part IV: A Model

We close with a model for coded peptide biosynthesis that incorporates these expanded amino acid site data, and also seems consistent with all that is known about RNA participation in translation. The result is an updated form of the DRT (Yarus 1998), or Direct RNA Template model.

In Fig. 7, panel A is a proposed primordial translation template, on which specific activated amino acids align for polymerization by binding directly to specific sites. Carboxyl activating groups (small dark circles) allow amino acid polymerization to yield a particular ordered (encoded) peptide. A monolithic DRT is shown, but it could also be composed of subunits, as RNA readily self-assembles into arrays (Jaeger et al. 2001).

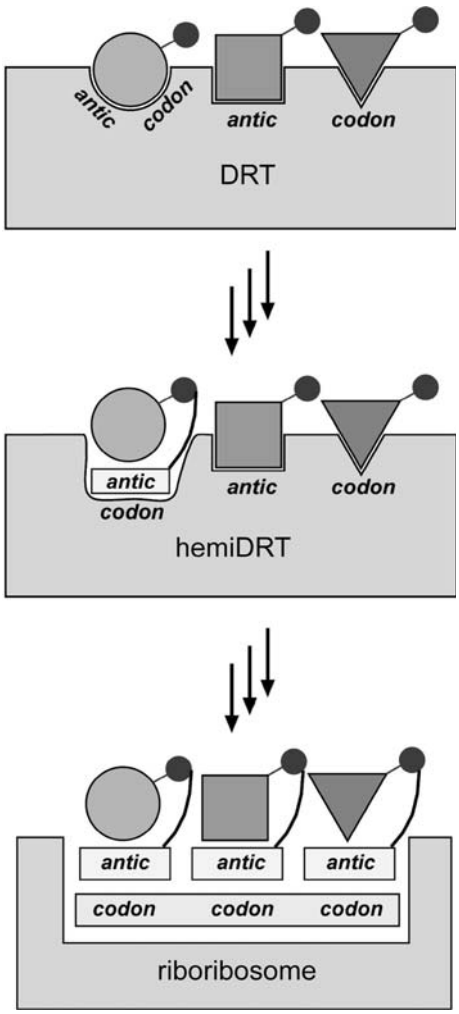


Fig. 7 A DRT (Direct RNA Template) model for the origin of coded translation. *antic* Anticodon, *hemiDRT* partially direct template, *riboribosome* RNA that hosts multiple encoded aminoacid polymerizations via codon–anticodon base pairing. Intermediate-sized shaded *circles*, *squares*, and *triangles* are particular amino acids, which are carboxyl-activated by a good leaving group (*small dark circles*). In the preferred form of DRT, activation in panel A is via esterification to ribose, adenosine, or AMP

Figure 7, panel B shows a possible evolving DRT system, which takes a step toward modern indirect coding by employing an RNA (tRNA-like) intermediate. The aa–RNA is elaborated from the original ribose or nucleotide activating group by incorporating a fragment of the DRT that includes the anticodon. This step would likely be first taken by amino acids whose sites have an unanticipated property emphasized by this present work; they incorporate both codons and anticodons at improbably high frequencies. Though Table 1 presents the aggregate data summarized over many sites, individual sequences like the arg #2 motif of Connell (Connell et al. 1993) do contain a complementary arginine CGG/CCG codon/anticodon pair. Thus the required informational materials to separate the

coding and anticoding functions preexist together. In panel B, the "escape" from initial binding function into nucleic acid coding function predicted in the "escaped triplet hypothesis" has occurred (Yarus et al. 2005). Escape can occur simply, without having to first invent a base-pairing RNA template (mRNA), as illustrated in panel B. It seems plausible, as (Szathmáry 1999) has suggested, that escape could also be selected for functional reasons independent of, and in addition to coding.

Figure 7, panel C illustrates the later transition to a uniform version of modern indirect coding. There is now a separate primitive mRNA, as well as activated forms of all amino acids associated with their anticodons (primitive aa-tRNAs). The larger part of the RNA holds these reactants and may have taken on other RNA functions, for example, that of peptidyl transferase (Nissen et al. 2000), to become a riboribosome.

There are substantial arguments in favor of updated DRT.

1. All chemistry that would be required for the Fig. 7 pathway, not just coding, is already known to be within the repertoire of small RNAs. Active RNAs that perform all translational reactions, or models of them, have been selected or are known (Yarus 2001). However and in particular, new data presented in this chapter on amino acid binding sites are consistent with and congenial to the DRT. A particular example is the definition of frequent, simple single-ended amino acid binding sites, whose focus on affinity for side chain atoms frees α-carbon substituents for the posited peptide forming reactions (Fig. 7).
2. The proposed DRT system of panel A suggests an exceedingly simple start point for the appearance of a primitive coding system, which would be of advantage in a primordial, barely controlled environment. Only two reactants, activated amino acids and the DRT itself, are initially required for useful translation.
3. The DRT model shows how accurately coded peptides of some complexity can appear before and without translocation, the most complex activity occurring on modern ribosomes.
4. Amino acid binding data and the consequent DRT model include an unexpected fraction of the modern coding apparatus even at the dawn of coding. The potential antecedents of the mRNA (codons), the tRNA (anticodons), and ribosome (the DRT) all exist in binding sites from the outset, facilitating evolution toward a modern system. The resulting transition to aa–RNA-mediated translation (panels 7 B & C) is plausibly selected because it enables all amino acids to be encoded without special chemistries to deal with unique RNA binding structures for every amino acid (see Part II, above). Using aminoacyl-RNAs instead of direct amino acid affinity, a single optimized molecular translation protocol can evolve for all amino acid side chains.

RNAs observably bind many (and may bind nearly all) biological amino acids with a simple chemical logic, approximated here in the polar profile. Binding is sufficiently strong and specific that it is easy to envision both easy initiation of coded peptide synthesis, as well as a continuous and plausible route therefrom, leading to modern translation. The success of experiments linking RNA chemistry to coding offers Bayesian support for the RNA world (Yarus 2001) and for the evolution of translation therein (Yarus et al. 2005). A salient problem now appears to be envisioning and testing an RNA- or RNA plus peptide-mediated fission of the DRT, as is required for the DRT progression of Fig. 7.

Acknowledgments Many thanks to National Institutes of Health R01 GM 48080 to MY and its Bioinformatics supplement, and to the NASA University of Colorado Astrobiology Institute NCC2-1052 to MY, and to NASA Astrobiology Grant EXOB07-0094 to RK, which supported parts of this work.

References

Blount KF, Wang JX, Lim J, Sudarsan N, Breaker RR (2007) Antibacterial lysine analogs that target lysine riboswitches. Nat Chem Biol 3:44–49

Caporaso JG, Yarus M, Knight R (2005) Error minimization and coding triplet/binding site associations are independent features of the canonical genetic code. J Mol Evol 61:597–607

Carothers JM, Oestreich SC, Davis JH, Szostak JW (2004) Informational complexity and functional activity of RNA structures. J Am Chem Soc 126:5130–5137

Carothers JM, Davis JH, Chou JJ, Szostak JW (2006) Solution structure of an informationally complex high-affinity RNA aptamer to GTP. Rna 12:567–579

Chen X, Li N, Ellington AD (2007) Ribozyme catalysis of metabolism in the RNA world. Chem Biodivers 4:633–655

Connell GJ, Illangsekare M, Yarus M (1993) Three small ribooligonucleotides with specific arginine sites. Biochemistry 32:5497–5502

Crick FHC (1968) The origin of the genetic code. J Mol Biol 38:367–379

Di Giulio M (1999) The coevolution theory of the origin of the genetic code. J Mol Evol 48:253–255

Dougherty DA (1996) Cation-pi interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp. Science 271:163–168

Dougherty DA (2007) Cation-pi interactions involving aromatic amino acids. J Nutr 137:1504S–1508S discussion 1516S-1517S

Ellington AD, Khrapov M, Shaw CA (2000) The scene of a frozen accident. RNA 6:485–498
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
Fitch WM, Upper K (1987) The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. Cold Spring Harb Symp Quant Biol 52:759–767
Freeland SJ, Hurst LD (1998) The genetic code is one in a million. J Mol Evol 47:238–248
Freeland SJ, Wu T, Keulmann N (2003) The case for an error minimizing standard genetic code. Orig Life Evol Biosph 33:457–477
Garst AD, Heroux A, Rambo RP, Batey RT (2008) Crystal structure of the lysine riboswitch regulatory mRNA element. J Biol Chem 283:22347–22351
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 SD, Rambo RP, Van Tyne D, Batey RT (2008) Structure of the SAM-II riboswitch bound to S-adenosylmethionine. Nat Struct Mol Biol 15:177–182
Hermann T, Patel DJ (2000) Adaptive recognition by nucleic acid aptamers. Science 287:820–825
Illangasekare M, Yarus M (2002) Phenylalanine-binding RNAs and genetic code evolution. J Mol Evol 54:298–311
Jaeger L, Westhof E, Leontis NB (2001) TectoRNA: modular assembly units for the construction of RNA nano-objects. Nucleic Acids Res 29:455–463
Knight RD, Landweber LF (1998) Rhyme or reason: RNA-arginine interactions and the genetic code. Chem Biol 5:R215–R220
Knight RD, Landweber LF (2000) Guilt by association: the arginine case revisited. RNA 6:499–510
Koonin EV, Novozhilov AS (2008) Origin and evolution of the genetic code: the universal enigma. IUBMB Life 61:99–111
Legiewicz M, Lozupone C, Knight R, Yarus M (2005) Size, constant sequences, and optimal selection. RNA 11:1701–1709
Lim J, Winkler WC, Nakamura S, Scott V, Breaker RR (2006) Molecular-recognition characteristics of SAM-binding riboswitches. Angew Chem Int Ed Engl 45:964–968
Lozupone C, Changayil S, Majerfeld I, Yarus M (2003) Selection of the simplest RNA that binds isoleucine. RNA 9:1315–1322
Lu C, Smith AM, Fuchs RT, Ding F, Rajashankar K, Henkin TM, Ke A (2008) Crystal structures of the SAM-III/S(MK) riboswitch reveal the SAM-dependent translation inhibition mechanism. Nat Struct Mol Biol 15:1076–1083
Majerfeld I, Yarus M (1994) An RNA pocket for an aliphatic hydrophobe. Nat Struct Biol 1:287–292
Majerfeld I, Yarus M (1998) Isoleucine: RNA sites with essential coding sequences. RNA 4:471–478
Majerfeld I, Yarus M (2005) A diminutive and specific RNA binding site for L-tryptophan. Nucleic Acids Res 33:5482–5493
Majerfeld I, Puthenvedu D, Yarus M (2005) RNA affinity for molecular L-histidine; genetic code origins. J Mol Evol 61:226–235
Mannironi C, Scerch C, Fruscoloni P, Tocchini-Valentini GP (2000) Molecular recognition of amino acids by RNA aptamers: the evolution into an L-tyrosine binder of a dopamine-binding RNA motif. RNA 6:520–527
Montange RK, Batey RT (2006) Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. Nature 441:1172–1175
Nagai K (1996) RNA–protein complexes. Curr Opin Struct Biol 6:53–61
Nissen P, Hansen J, Ban N, Moore PB, Steitz TA (2000) The structural basis of ribosome activity in peptide bond synthesis. Science 289:920–930
Orgel LE (1968) Evolution of the genetic apparatus. J Mol Biol 38:381–393
Pugilisi JD, Chen L, Frankel AD, Williamson JR (1993) Role of RNA structure in arginine recognition of TAR RNA. Proc Natl Acad Sci USA 90:3680–3684
Serganov A, Huang L, Patel DJ (2008) Structural insights into amino acid binding and gene control by a lysine riboswitch. Nature 455:1263–1267
Sheppard K, Yuan J, Hohn MJ, Jester B, Devine KM, Soll D (2008) From one amino acid to another: tRNA-dependent amino acid biosynthesis. Nucleic Acids Res 36:1813–1825
Sokal R, Rohlf F (1995) Biometry: the principles and practice of statistics in biological research. Freeman & Co., New York
Sudarsan N, Wickiser JK, Nakamura S, Ebert MS, Breaker RR (2003) An mRNA structure in bacteria that controls gene expression by binding lysine. Genes Dev 17:2688–2697
Szathmáry E (1999) The origin of the genetic code: amino acids as cofactors in an RNA world. Trends Genet 15:223–229
Tao J, Frankel AD (1996) Arginine-binding RNAs resembling TAR identified by in vitro selection. Biochemistry 35:2229–2238
Wang X, McLachlan J, Zamore PD, Hall TM (2002) Modular recognition of RNA by a human pumilio-homology domain. Cell 110:501–512
Wang JX, Lee ER, Morales DR, Lim J, Breaker RR (2008) Riboswitches that sense S-adenosylhomocysteine and activate genes involved in coenzyme recycling. Mol Cell 29:691–702
Woese CR (1967) The genetic code: the molecular basis for genetic expression. Harper & Row, New York
Wong JT-F (1975) A co-evolution theory of the genetic code. Proc Natl Acad Sci USA 72:1909–1912
Wong JT-F (1981) Coevolution of genetic code and amino acid biosynthesis. Trends Biochem Sci 6:33–36
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
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 (2001) On translation by RNAs alone. Cold Spring Harb Symp Quant Biol 66:207–215
Yarus M, Christian EL (1989) Genetic code origins. Nature 342:349–350
Yarus M, Majerfield I (1992) Co-optimization of ribozyme substrate stacking and L-arginine binding. J Mol Biol 225:945–949
Yarus M, Caporaso JG, Knight R (2005) Origins of the genetic code: the escaped triplet theory. Annu Rev Biochem 74:179–198
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