J Mol Evol (2005) 61:226–235
DOI: 10.1007/s00239-004-0360-9

JOURNAL OF MOLECULAR EVOLUTION

RNA Affinity for Molecular L-Histidine; Genetic Code Origins

Irene Majerfeld, Deepa Puthenvedu, Michael Yarus

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

Received: 15 December 2004 / Accepted: 25 February 2005 [*Reviewing Editor:* Niles Lehman]

Abstract. Selection for affinity for free histidine yields a single RNA aptamer, which was isolated 54 times independently. This RNA is highly specific for the side chain and binds protonated L-histidine with 10^2–10^3-fold stereoselectivity and a dissociation constant (K_D) of 8–54 μM in different isolates. These histidine-binding RNAs have a common internal loop–hairpin loop structure, based on a conserved $RAAGUGGGKKN_{0–36}$ $AUGUN_{0–2}AGKAACAG$ sequence. Notably, the repetitively isolated sequence contains two histidine anticodons, both implicated by conservation and chemical data in amino acid affinity. This site is probably the simplest structure that can meet our histidine affinity selection, which strengthens experimental support for a ''stereochemical'' origin of the genetic code.

Key words: Selection — SELEX — Amino acid — Coding — Triplet

Introduction

The 20 standard biological amino acids are interesting ligands for RNA. They present chemically varied surfaces, and RNA's response to these chemical challenges should be relevant to the intrinsic possibilities of RNA–peptide interfaces. Furthermore, stereochemical theories of the origins of the genetic code (Woese et al. 1996) propose that chemical affinities between codons and/or anticodons and amino acids determined at least some codon assignments. Such affinities between RNA and amino acids are known to be real and varied. The first was found in the group I self-splicing RNA, where arginine acts as a competitive inhibitor for the guanosine splicing co-factor. Binding is therefore within the active site (Yarus 1988). Because this binding site depends on conserved arginine codons (Yarus and Christian 1989), the group I active center appears to be a ''molecular fossil'' of codon assignment to arginine (Yarus et al. 2005), possibly dating from the RNA world.

The example of the group I active center also suggested that triplet–amino acid chemical association might be best observed within longer, more stable oligonucleotide foldings. If these findings generalize, similar associations between coding sequences and amino acids might be recaptured within the amino acid binding motifs of newly selected RNA aptamers isolated using selection–amplification or SELEX.

A number of such aptamers have been isolated and their binding sites identified by conservation, NMR, and/or chemical probing. These include an aptamer for tryptophan-agarose (Famulok and Szostak 1992), several aptamers for arginine (Burgstaller et al. 1995; Connell et al. 1993; Connell and Yarus 1994; Famulok 1994; Geiger et al. 1996; Yang et al. 1996), valine (Majerfeld and Yarus 1994), isoleucine (Majerfeld and Yarus 1998), tyrosine (Mannironi et al. 2000), phenylalanine (Illangasekare and Yarus 2002), glutamine (Tocchini-Valentini, personal communication; Yarus et al. 2005), and leucine (Majerfeld and Yarus, unpublished; Yarus et al. 2005), as well as histidine

Correspondence to: Michael Yarus; *email:* yarus@colorado.edu

and tryptophan (this work; Yarus et al. 2005). If there is association between triplets and their cognate amino acids, coding sequences will appear in binding sites more frequently than chance suggests. Statistical analysis (Knight and Landweber 1998; Yarus et al. 2005) compares the abundance of coding triples within the aptamer binding site to their abundance in the nonsite sequences from the same molecule. Data exists for eight amino acids, and significant associations were found for both codons and anticodons, In fact, the overall probability of the null hypothesis—that there is no association between these eight kinds of amino acid binding sites and cognate coding triplets—now stands at 5.4×10^{-11}. This is exceedingly unfavorable to the null hypothesis and in support of a robust relationship between coding triplets and amino acid affinity, especially given that the data include both strongly negative and positive cases.

In order to increase the generality of this test of amino acid binding sites, we have selected, using a multitarget column affinity protocol, aptamers to histidine, tryptophan, glutamine, and valine. This selection yielded many independently occurring sites for histidine and tryptophan. Here we report the molecular characterization of the histidine sites. Results for tryptophan aptamers will be reported elsewhere.

A precedent for a histidine-binding nucleic acid can be found in a class of *in vitro* selected DNAzymes with RNA cleavage activity dependent on histidine (Roth and Breaker 1998). Activity is pH dependent in a manner that implicates the imidazole group of histidine in the chemical step of the cleavage reaction. A cofactor binding site with an apparent K_D of ~ 25 m*M* was demonstrated for the variant HD1, Like the aptamers to be described below, these DNAzymes strongly discriminate against the D-isomer, showing thousandfold lower activity.

Methods

Preparation of the Selection Matrix

Fmoc-protected glutamine, histidine, tryptophan, and valine were coupled separately to EAH Sepharose 4B (Amersham) via their carboxyl groups, as described before (Illangasekare and Yarus 2002) except that pentaflurophenyl ester (OPfp) preactivated Fmoc derivatives were used. These preparations, with final amino acid concentrations of approximately 1.5 to 1.9 m*M*, were mixed to produce an affinity matrix of 0.4 m*M* each amino acid. Counterselection matrix was prepared by reaction of the same Sepharose with a 30-fold excess of acetic anhydride with respect to resin amino groups.

Selection Procedure

Approximately 10^{14} DNA sequences were amplified and transcribed with T7 RNA polymerase to obtain the starting RNA pool. Primer regions flanking the 70 nucleotide long randomized tract in the initial DNA avoided codons and anticodons for the selective amino acids. The initial DNA sequence was taatacgactcactatagg gatcctaagctctatcgg N(70) aaagcggcctagcgatcga where underlined nucleotides indicate the T7 promoter.

Selection buffer was 50 m*M* Hepes (pH 7.0), 250 m*M*, NaCl, 5 m*M* each $CaCl_2$ and $MgCl_2$, and 1 m*M* glycine (but see below). Except for the first round of selection all cycles were preceded by counterselection on a 0.2-ml acetylated column. After applying the internally ^{32}P-labeled RNA (1 nmol for cycle 1), 0,4-ml selection columns were washed with 4.5 bed volumes of selection buffer and the remaining bound RNA was eluted with 1.6 ml of the same buffer containing, in addition, the selection amino acids. The resulting pools were reversed transcribed, the DNA was amplified and transcribed into RNA (Ciesiolka et al. 1996) for the following cycle. Mutagenic PCR (Cadwell and Joyce 1994; Wilson and Keefe 2000), resulting in approximately 1 mutation per 100-mer per PCR procedure, was used after cycles 3, 4, and 5.

Two parallel selections were carried out that differed only in the concentration of affinity eluants. In selection I, the bound RNA in cycle 1 was eluted with a mixture of 1 m*M* each selection amino acid. Then the concentrations were decreased to 0.8, 0.6, 0.4 and 0.2 m*M* each in successive rounds and kept at the latter concentration for the rest of the selection. The glycine concentration in selection I buffers was decreased to match that of the selection amino acids. In selection II, the eluants' concentrations were 1 m*M* each throughout. The final pools were cloned (Novagen PT7 Blue-3 Perfectly Blunt Cloning Kit) and individual clones were sequenced. The resulting pools of histidine and tryptophan RNA aptamers had no recurring sequences that were exclusive to selection I or selection II. Eight more sequential selection cycles failed to produce specific glutamine and valine specific affinity responsive to free ligands.

Chemical Probing

Chemical modifications were done essentially as described by Kroll and Carbon (1989). DMS (dimethyl sulfate) reactions were done in selection conditions except the Hepes concentration was increased to 100 m*M*. For CMCT (l-cyclohexyl-3-[2-morpholinoethl] carbodiimide) reactions the buffer was 50 m*M* sodium borate, pH 8.0; otherwise reactions were as for the DMS reaction. For protection experiments 10 to 20 pmol of folded RNA (5′ ^{32}P-labeled for G detection) was modified in the absence or presence of 800 μM and 8 m*M* histidine. Modified positions were detected by primer extension with AMV reverse transcriptase (A and C) or by $NaBH_4$ reduction followed by aniline cleavage (G) on denaturing 10% polyacrylamide gels. For interference experiments, folded RNA was modified with DMS or CMCT, the buffer was exchanged by passage through a P6 column, and the modified RNA was fractionated by passage through a histidine–Sepharose column into flowthrough and specifically eluted fractions. Elution employed 1 m*M* histidine after a 4.5-column volume wash. Modified positions with specific relation to affinity were detected by comparison of primer extensions, using PAGE.

Results

Selection

Histidine aptamers were selected from an initial pool of approximately 10^{14} unique 70 nucleotide randomized sequences in two simultaneous, multitarget selections for the amino acids glutamine, histidine, tryptophan, and valine (Morris et al. 1998). In cycle 6, elution peaks were observed (5 and 12% of applied

RNA for selections I and II). RNA from these pools responded only to histidine. Pooled RNA from cycle 6 was cycled once more. After applying the RNA, the selection column was washed with the three other selection amino acids before eluting with histidine. Both cycle 6 and cycle 7 pools were cloned and sequenced to get the aptamers for this work. The selection was continued for the three remaining aminoacids and a pool of tryptophan aptamers was obtained in cycle 8 after prewashing with the other three amino acids.

The selection was continued for 8 more cycles but no response to glutamine or valine was detected. For cycles 11 to 14 the RNA applied to the selection column was washed with histidine and tryptophan before eluting with glutamine and valine; the last two cycles were run in separate columns, washing with the three other amino acids before eluting with glutamine or valine. Although no bulk elution peaks were observed, cycle 16 RNA from both pools was cloned and sequenced. We obtained 25 sequences from the valine selection pool and 16 from the glutamine pool. Since some sequences appeared in both selections, we will describe the outcome of the selections for the combined pools. Twenty-nine sequences were enriched into seven families (sequences from a single parent) of 2 to 9 clones; the remaining 12 were single sequences, indicating that selection was complete because pool sequence variety had been depleted.

Examination of these sequences revealed 14 isolations of the histidine motif to be described below, often including nucleotides that are less frequently used in isolates from cycles 6 and 7. There was one example of the dominant tryptophan motif and four nonidentifiable sequences. Three of the sequences containing the histidine motif were also present in the earlier histidine-specific pool. Seven of the enriched sequences were assayed for binding to and elution from the selection matrix. In general they did not bind strongly to the column and those carrying the histidine motif responded poorly to the free amino acid. Some appeared to have additional affinity for the selection matrix itself because they were eluted slowly and trailed off the selective column. Those characteristics help explain their low level persistence in successive selection pools. Histidine-binding RNAs from cycle 16 were not included in the analysis reported below. None of the assayed sequences responded to glutamine or valine. Since glutamine (Tocchini-Valentini, personal communication; Yarus et al. 2005) and valine (Majerfeld and Yarus 1994) RNA aptamers have previously been isolated, this selection may have been too stringent for isolation of aptamers that bind difficult targets.

We sequenced 134 clones, 46 from cycle 6, the rest from cycle 7, and grouped them according to sequence similarity (PILEUP; Wisconsin Package Version 10.1). The complete list of sequences is available online as supplementary material and is summarized in Figure 1, aligned on completely conserved nucleotides at the 3′ end of the motif. The sequences in Figure 1 represent 83% of the pool (111 sequences) and include all those that fulfill the selection requirement of binding histidine. Fourteen sequences occurred in families of 2 to 20 clones each; for families, only 1 member is shown. There were also 40 unique sequences. All sequences share the same two conserved segments, (shaded nucleotides are at least 62% conserved): AAGUGGG and AUGUN (0–2) AGUAACAG, the second always 3′ to the first. We refer to these as module 1 (5′) and module 2, respectively. The nucleotide spacing between the two modules varies between 0 nucleotides (Fig. 1; sequence His 1bd) and 36 nucleotides (sequence His 235b). Both conserved elements are involved in folding the binding site.

The remaining 17% of the pool, not represented in Figure 1, consists of unique sequences. Since four of these were tested and failed to bind free histidine, the group was not further analyzed, and in what follows the frequencies that we quote are in reference to the 111 sequences that contain the conserved site. Summarizing, one single motif formed by two independent modules, dominated this selection, and it was isolated independently 54 times within the 134 sequences of the pool. Because it was isolated in only one sequence permutation despite many independent origins, it likely forms one motif with a required sequence disposition around conserved loop sequences, so that its elements cannot be successfully permuted. Alternatively put, the hypothesis that site sequences can be freely permuted has a probability of 6×10^{-17} at this point, thereby vanishingly small.

Amino Acid Binding by the Isolates

Clones from the various families and single sequences represented in Figure 1 bind well to a histidine–Sepharose column and elute readily when histidine is added to the column buffer. Figure 2A illustrates this observation for His 225, a representative of a seven-member family (Fig. 1). A fraction of the RNA, probably aggregating or failing to fold properly, did not bind to the column but bound RNA radioactivity was eluted quantitatively by the addition of 1 m*M* histidine. That is, none remained to be removed by a high-salt-EDTA wash. Furthermore, as shown in Figures 2B and C, the binding is specific for the histidine side chain since other amino acids of similar size or charge, like glutamine, lysine, tryptophan, and arginine, failed to elute bound RNA when applied at a concentration that is 100 times greater than the K_D for histidine for these clones (see below). A contri-

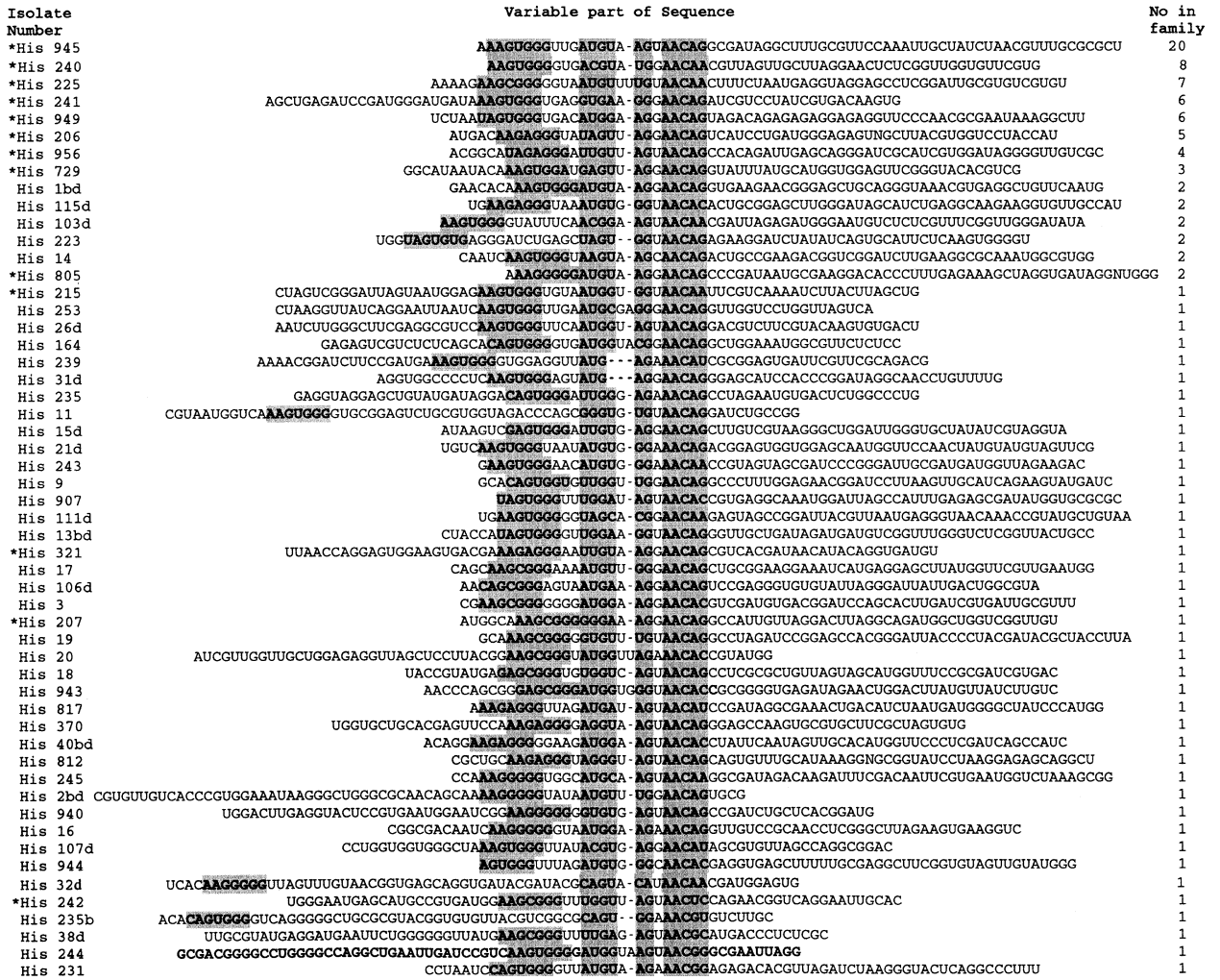

Fig. 1. Composition of the selected pool. Alignment was based on the 3′ conserved segment. One representative from each family and the number of sequences isolated are shown. Shaded nucleotides are at least 62% conserved. Sequences labeled "d" were from the cycle 6 pool; the remaining, from cycle 7. Sequences marked with an asterisk were used for chemical probing to identify the binding site nucleotides. Unique sequences that did not contain these conserved segments are not shown but are available online.

bution of the specific chemical form side chain to binding is also suggested by the failure of charged methyl-amine to elute His 945 RNA (Fig. 2C).

We estimated dissociation constants (K_D) for free L- and D-histidine by isocratic elution from histidine-Sepharose for some isolates (Ciesiolka et al. 1996). For three RNAs we obtained dissociation constants from the variation of DMS protection for the conserved nucleotides AACA in Module 2 as the histidine concentration varied (Welch et al. 1995). The K_D we obtained for L-histidine varied between 8 and 54 μ*M* (Table 1). There was good agreement between data obtained by affinity chromatography and by DMS protection, the second method being more transparent because no matrix is involved. K_D for the D-isomer varied between 3 and 9.5 m*M*. Greater affinity for L-histidine, used for selection, over D-histidine ranged from 130-fold for His 225 to 860-fold for His 945. Since our procedures for K_D determination both directly determine the affinity for free amino acid (Ciesolka et al. 1996), observed differences may be associated with sequence variation within the motif. Sequence comparison suggests that the higher stereospecificity of His 945 might be associated with the second (3′) module majority sequence AACAG versus the sequence AACAA present in His 225 and His 240. This in turn is most simply explained if these sequences are in close contact with bound ligand, as suggested by chemical protection.

Figure 3 shows binding profiles for His 945 RNA at different pH levels. At pH 5.5, 85% of the applied RNA eluted specifically from the column, this proportion was reduced as the pH was increased, more significantly above pH 7.0, dropping to approximately 50% at pH 7.25. Because affinity increases as pH drops near the known pK_a of histidine, these observations probably mean that the protonated

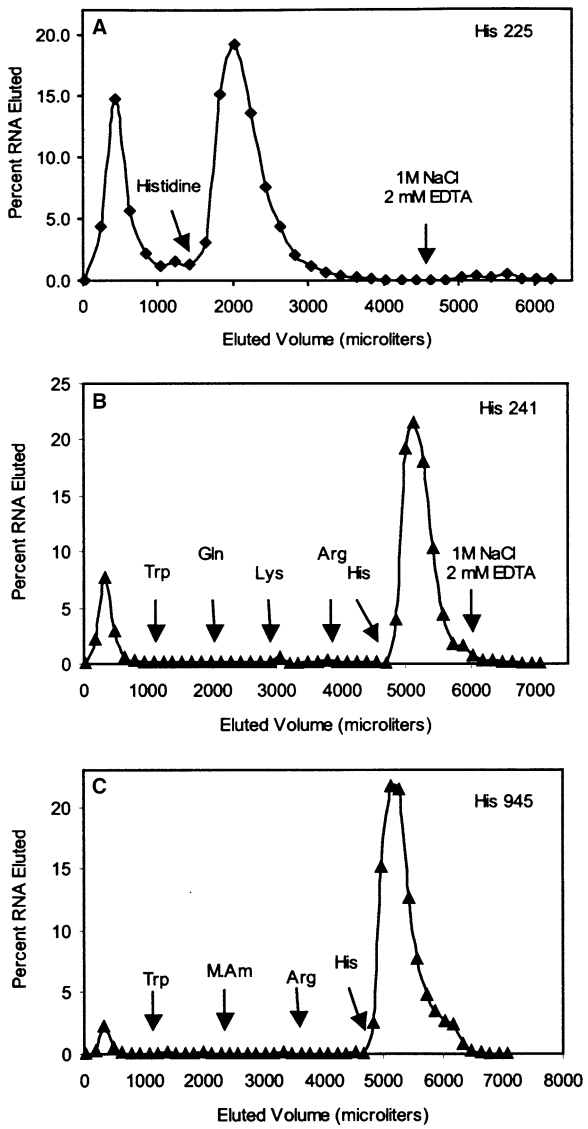

Fig. 2. Affinity chromatography of RNA on 0.6 m*M* histidine–Sepharose. Thirty to fifty picomoles of internally ^{32}P-labeled, folded RNA was applied to a 0.3-ml column pre-equilibrated with selection buffer. Eluants were added at a 1 m*M* concentration. **A** Binding and histidine elution of clone His 225. **B**, **C** Specificity of ligand binding. M.Am, methyl amine.

form of histidine is bound by RNA. The shift to an apparent pK_a for binding around 7 may indicate that histidine is more easily protonated when protonation is coupled to binding to an RNA site. All other experiments reported here were done at pH 7, the pH at which the selection was performed.

Identification and Characteristics of the Binding Site

A purpose of this work is to identify the binding site and to determine the content of coding triplets, codons, and anticodons inside and (as a control) outside the active sequences. A statistical overabundance of codons and/or anticodons within the binding site would suggest a chemical affinity between an amino acid and RNA aptamers containing its coding triplet, an association that supports a stereochemical origin for the genetic code (Yarus et al. 2005).

Nucleotides relevant to histidine binding were identified by chemical probing with DMS and CMCT. Ligand-induced protections from DMS modification identified nucleotides with reactivities affected by histidine binding (Krol and Carbon 1989). Modifications at the N-1 of A and the N-3 of C were detected by reverse transcription, reaction at the N-7 of G by chain scission. Thirteen independently arising sequences were examined with either one or both assays. In addition, interferences to column binding by modified RNA were determined for four of those sequences. For the interference experiments, folded RNA was first modified with DMS (N-1 of G and N-3 of C) or with CMCT (N-1 of G and N-3 of U), then applied to the histidine column to separate binding and nonbinding fractions. Molecules with modifications at positions essential for binding should be enriched in the flowthrough. Results are presented in detail for His 945, which represents the most frequent recovered sequence in the selected pool and also contains the most frequently occurring nucleotides at positions within the conserved sequences where conservation is not complete (Figs. 4 and 5). Figure 6A summarizes the data for all tested clones.

Figure 4 shows the most likely structure calculated by Bayesfold (Knight et al. 2004) from the variation in 20 aligned recurrences of His 945. This predicted fold utilized not only the 20 aligned sequences but, in addition, DMS and CMCT chemical accessibility data for unpaired residues. Results from chemical probing are summarized in Figure 4, with the data from protection and interference experiments shown in Figure 5. Protections from DMS modification are shown in Figures 5A and B; modification-interference by DMS and CMCT products, in Figure 5C. Sensitive nucleotides are concentrated in the 5′ half of the molecule, covering an asymmetrical loop, a terminal loop, and a short stem linking the two. The residues involved overlap the conserved segments highlighted in Figure 1—in mutual confirmation of each kind of data, a majority of conserved positions were also detected by a chemical assay for essential nucleotides.

The short sequence A40 A41 C42 A43 is only moderately accessible to DMS modification (Fig. 5A), and indeed it is predicted to take part in a short stem. However, interferences in this segment were pronounced (Fig. 5C) not only in His 945 but in all tested clones (see Fig. 6A). The following nucleotide, G44, is highly conserved, present in 75% of sequences (see below), and its reactivity is not affected by histidine binding. However, in three sequences tested, His 240, His 225, and His 215, in which this G is replaced by an A that can be modified effectively at

Table 1. Dissociation constants (K_D; μM) for histidine

	L-Histidine						
		DMS protection					D-Histidine
Isolate	Affinity chromat.	A	A	C	A	Average	Affinity chromat.
His 945	9/7	17		12		11 ± 2	9500
His 240	22/24	25	17	17	21	21 ± 2	2800
His 225	8/16					12	3000
His 206	48/52	55	62	55		54 ± 2	
His 956	9/12					11	
His 215	8/7					8	
His 321	17/28					22	
His 945 fragment	9/15					12	

Note. In columns headed Affinity Chromat., K_D's were determined by the two-column affinity procedure, eluting with and without isocratic ligand (Ciesiolka et al. 1996). Under DMS protection, K_D's reflect the progressive protection from chemical attack at increasing ligand concentrations, linearized and fit with a least-squares line (Welch et al. 1995). AACA is a highly conserved sequence within Module 2; values are given wherever data were sufficiently distinct from background densities. D-Histidine values presented without accompanying errors (±standard error of the mean) are averages of two experiments.

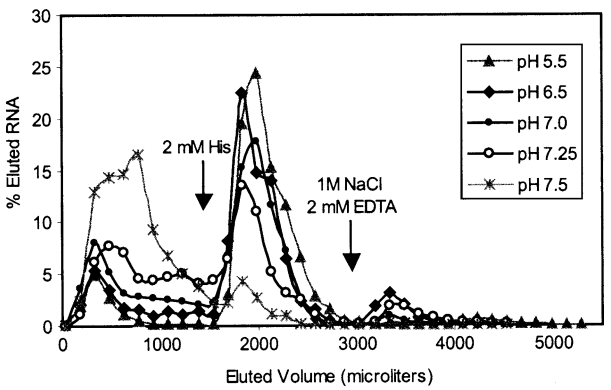

Fig. 3. Dependence of histidine affinity on pH. His 945 RNA was folded at the corresponding pH, applied to the column, and specifically eluted after a 5-column volume wash. Column buffers were prepared by adjusting Hepes to the desired pH; other constituents were unchanged.

the N-1 position, both protections and interferences were seen, suggesting that the Watson–Crick face of nucleotide 44 could be involved in constructing the histidine pocket (Fig. 6A). This reinforces the impression that they are close to the ligand, as surmised from the role of these nucleotides in the stereoselectivity of the binding site, noted above.

Figure 6A summarizes the results of chemical probing for all sequences tested, from families as well as single sequences (see Fig. 1). Some—His 956, His 207, His 215, and His 321—were tested only for protections by primer extension, so only sensitive As and Cs could be identified. His 805 was tested only for sensitive Gs. In addition, interferences were tested for His 945, His 240, His 215, and His 321. Results confirmed the data obtained for His 945—protections and interferences, within experimental error, not only consistently concentrated in the conserved segments, but gave a similar pattern of reactivities. These data therefore support the sequence conservations in suggesting that all sequences form a common site with a common affinity mechanism.

Figure 6B shows the consensus sequence and nucleotide conservation for the histidine motif. Module 1 has the consensus sequence AAGUGGG, with all nucleotides at least 62% conserved; among these, 3 positions are invariant and 2 are at least 96% conserved. Module 2, with the consensus sequence AUGUN(0–2)AGKAACAG, has 11 positions that are a least 62% conserved. Of these, 4 positions are invariant and 2 are at least 95% conserved.

Two histidine anticodons are present in the majority of sequences, GUG in module 1 and AUG in module 2. GUG is present in 62% of sequences and was isolated independently 28 times. AUG is present in 66% of sequences and was isolated 31 times. For both, chemical probing indicates involvement of triplet nucleotides in forming the binding site (Fig. 6A). The two flanking Gs in the GUG anticodon are invariant while the middle nucleotide can be varied without apparent detriment to binding (Fig. 1; compare K_D for His 945 and His 225). Neither is specificity affected. His 225, when bound to the histidine affinity column, is not eluted by tryptophan, glutamine, lysine, or arginine as exemplified in Figure 2B for isolate His 241. In spite of this flexibility, this U nucleotide is implicated in the binding site in the majority of the clones (Fig. 6A) by a consistent interference enhancement. Therefore, all three GUG nucleotides are amino acid site nucleotides.

The AUG triplet contains an invariant and consistently protected 3′ G in module 2. AUG is predominant in the pool, but its function may be replaced with other less evident structures because K_D is similar whether sequences contain AUG, as in His 945, or not, as in His 240 (Fig. 1). Specificity is

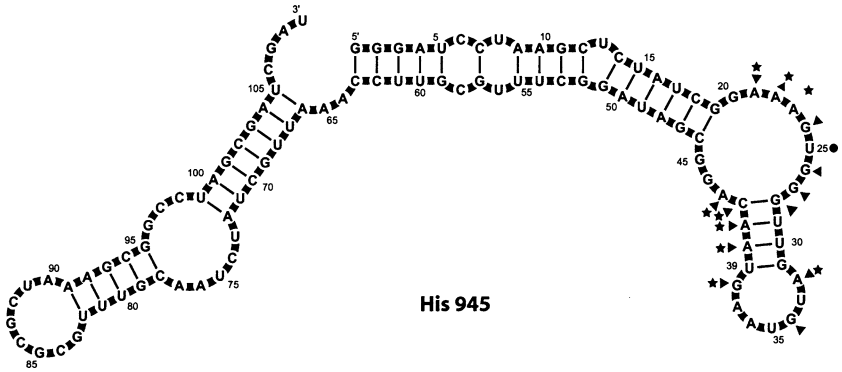

Fig. 4. Chemical probing of His 945 RNA. Results are summarized using the most probable secondary structure from Bayesfold, based on the variation in aligned sequences (Knight et al. 2004). (▼) Bases protected from DMS modification by histidine; (▲) bases with reactivities to DMS stimulated by histidine; (*) bases that when modified by DMS or CMCT interfere with binding to the affinity column; (●) bases that when modified enhance binding to the affinity column.

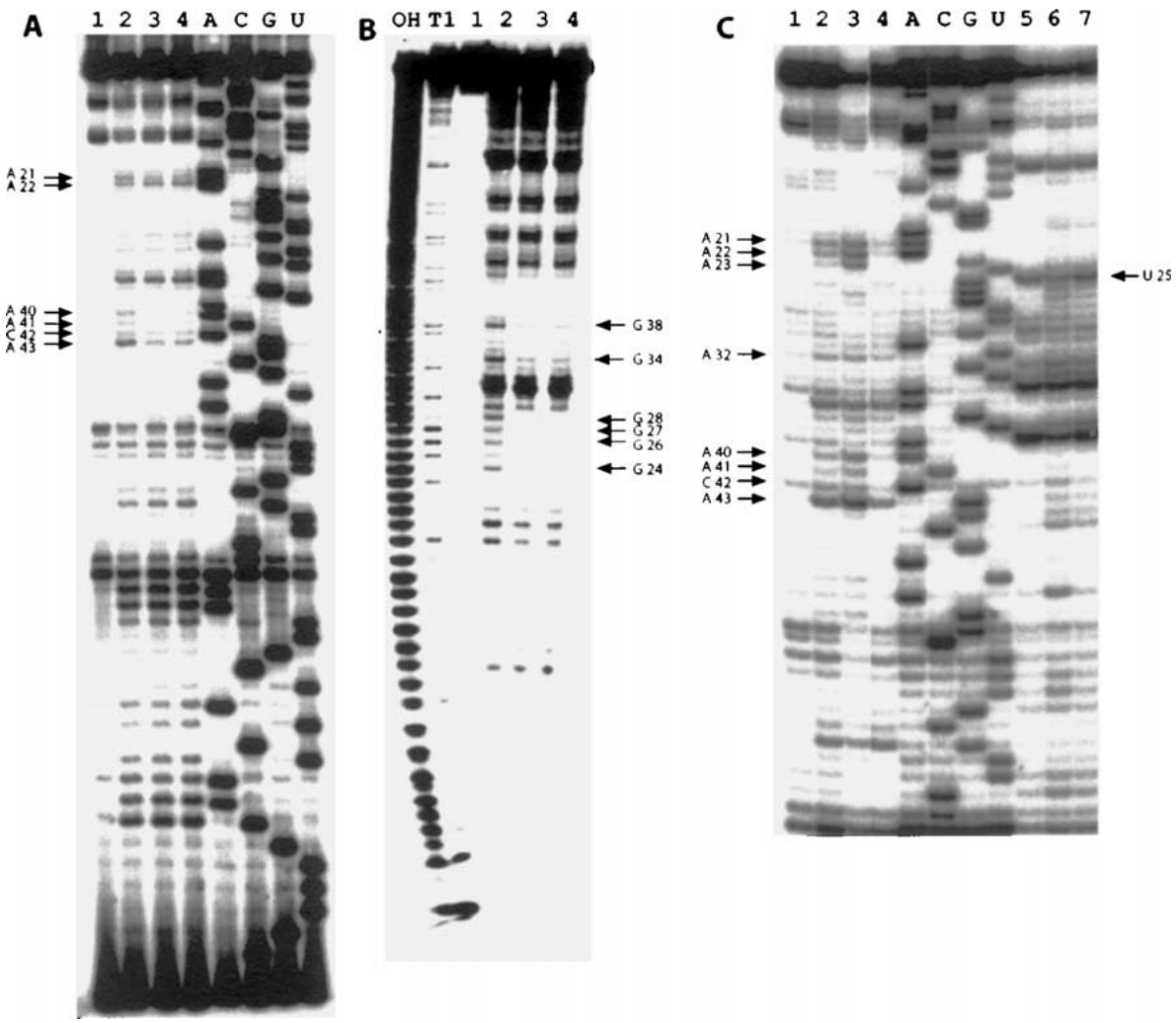

Fig. 5. Chemical probing of His 945 RNA. Positions with altered reactivities are indicated on the gels. **A** DMS modifications and footprinting. Lane 1 is a control reaction in the absence of DMS. Lane 2 represents a reaction in the absence of ligand. Lanes 3 and 4 are reactions in the presence of 0.8 and 8 m*M* histidine. A, C, G, and U are dideoxy sequencing lanes. **B** DMS modification and chain scission. OH indicates partial base hydrolysis; T1 indicates limited T1 RNase hydrolysis. Lane 1 is an untreated control. Lane 2 is a reaction in the absence of ligand. Lanes 3 and 4 are reactions in the presence of 0.8 and 8 m*M* histidine. **C** DMS (lanes 1–4) and CMCT (lanes 5–7) modification-interference. Lane 1 is untreated RNA. Lanes 2 and 5 contain unfractionated RNA. Lanes 3 and 6 are unbound fractions from affinity chromatography. Lanes 4 and 7 are bound and specifically eluted fractions. A, C, G, and U are sequencing lanes.

also grossly unaffected by this variation (Figs. 2B and C; compare His 945 to other sequences). Again, protection and interference identify the flanking nucleotides in the AUG as essential site nucleotides, and its 85% conservation suggests a role for the central U.

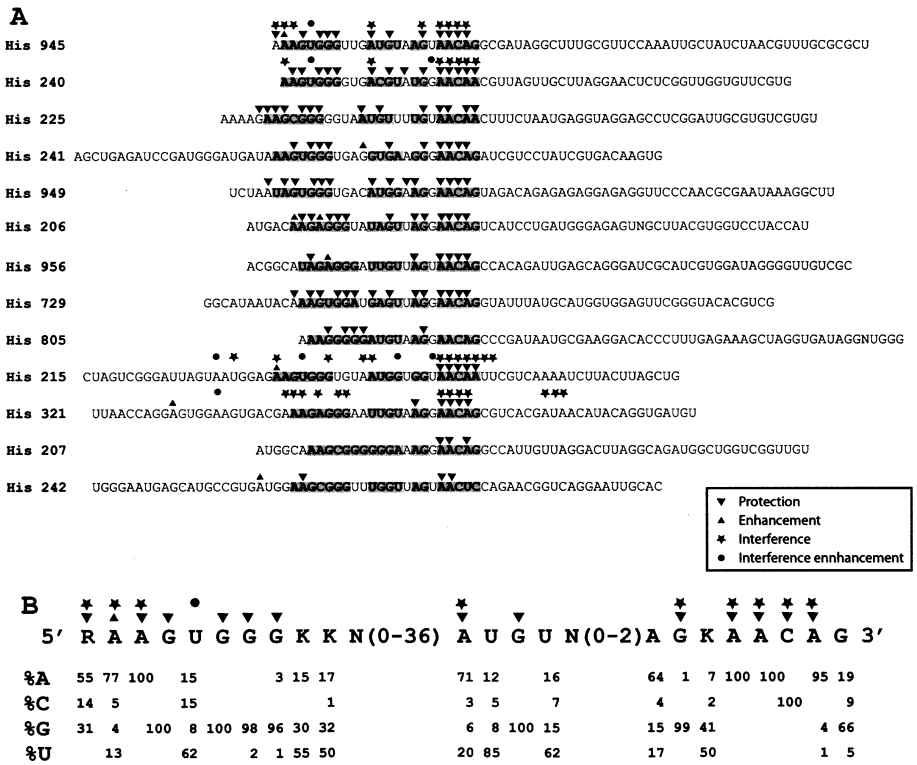

Fig. 6. Summary of chemical probing data for 13 clones. Positions with reactivities altered by histidine or at which modifications affect histidine binding are superposed on the variable region of the clone. Shaded nucleotides indicate conservation (at least 62%) as in Fig. 1. Sequences His 956, His 321, His 207, and His 242 were tested only for protections on A and C; His 805, only for protections on G. His 945, His 240, His 215, and His 321 were also probed for modifications that interfere with binding. **B** Consensus sequence and nucleotide variation within the histidine motif. The chemical probing data are consistent with those shown in A.

In summary, the histidine motif usually extends through 22–24 nucleotides, 11 positions being at least 95% conserved. The conserved nucleotides spread over an asymmetrical loop, a terminal loop, and a short stem linking the two. One or both histidine anticodons are present in 86% of the sequences. These nucleotides contribute to binding but substitutions in the middle nucleotide of GUG or the first or second nucleotide in AUG are occasionally observed.

The Minimal Active Sequence

We determined the 5′ and 3′ boundaries for minimal active molecules of His 945 by partial alkaline hydrolysis followed by binding and elution from the histidine column (not shown). These experiments suggested a 40-nucleotide-long minimal fragment which proved to have a 12 μ*M* K_D. This fragment includes all conserved nucleotides highlighted in Figure 1. A predicted secondary structure for the minimized RNA is shown in Fig. 7A with its binding and elution profiles. We introduced a number of mutations not predicted to affect the predicted folding to test the importance of nucleotides highlighted by conservation and chemical probing. For example, a change in module 1 from AAGUGGG to AGACAGG (Fig. 7B) or AAGCAGG (both "anticodon changes"; Fig. 7C) completely abolished binding. Equally detrimental were changes in the terminal loop from AUG to GCA (an anticodon sequence; Fig. 7D) or to ACA (Fig. 7E). Reversal of three of the four nucleotides in the short stem that links the two loops had the same effect (Fig. 7F). The best-tolerated change was that of AACAG to AACGG (Fig. 7G), which left some binding activity. In agreement with this result, four single sequences carrying this modification were found in the original pool (Fig. 1; last four sequences). One of these, His 244, was tested for binding and elution from histidine–Sepharose and its behavior was that of a typical aptamer as exemplified by His 225 in Figure 2A.

These data are consistent with the notion that both conserved modules, including both loops with their anticodons and the connecting stem, are parts of essential structures, as implied by their multiple independent isolation in the pool.

Discussion

A selection for L-histidine affinity at pH 7, where the imidazole sidechain would be partially protonated, has yielded a single predominant site for free, protonated histidine. This site, recovered independently from 54 different parental sequences in the initial pool, comprised 111 sequences of the 134 RNAs cloned and sequenced. Individual isolates show a mean K_D for L-histidine of 19 μ*M* (Table 1) and strong unselected stereoselectivity against D-his of ≈100- to 1000-fold.

Variation in the sequence families supports a structure in which the histidine site is formed by two

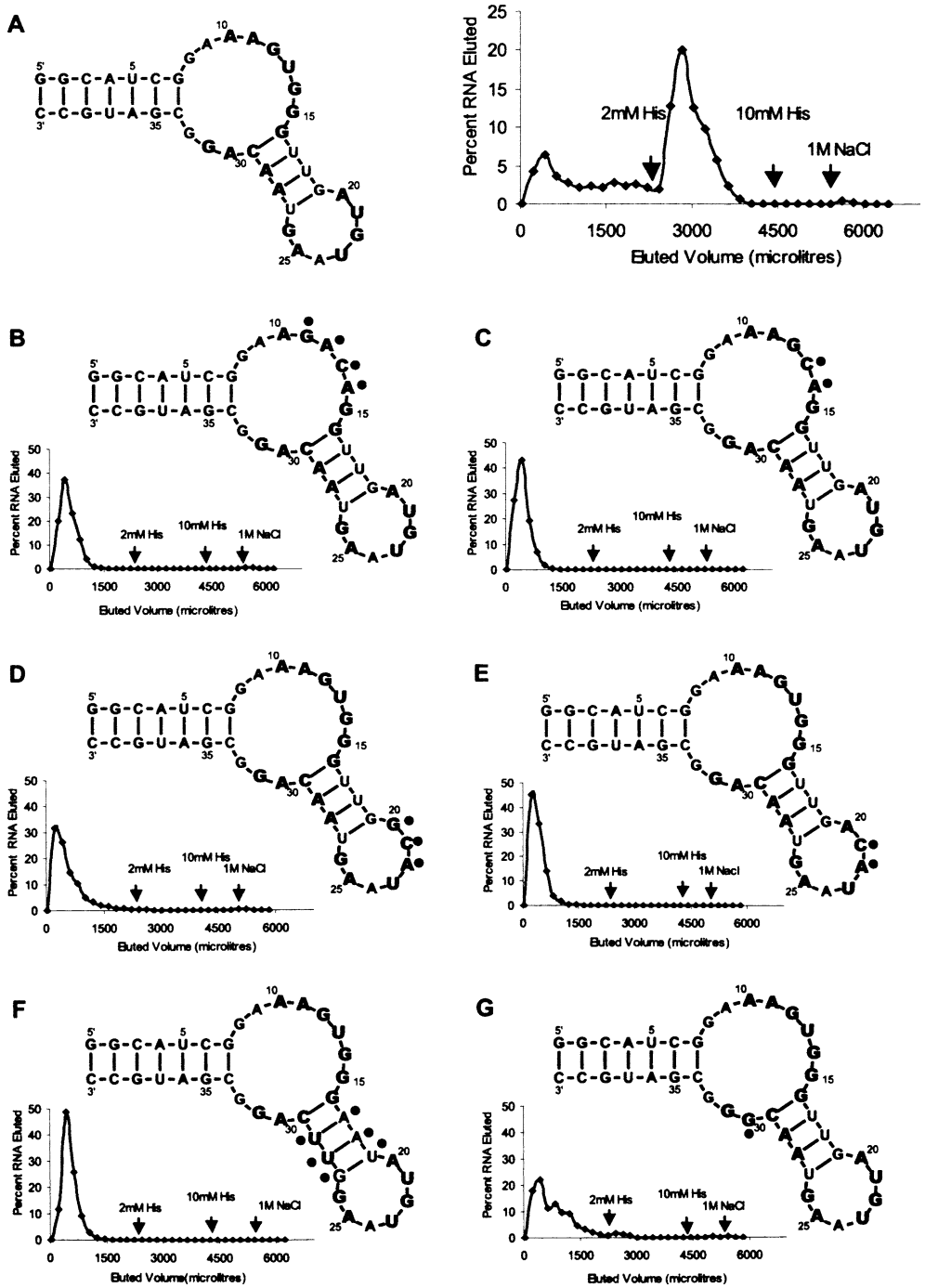

Fig. 7. Effect of mutations in conserved nucleotides of a His 945 fragment. **A** His 945–40 predicted secondary structure and column chromatography on histidine sepharose. B–G Variant fragments and their binding and elution behavior. (●) Changed position. Additions to the column buffer are indicated by arrows. Boldface nucleotides are >62% conserved in the pool.

conserved sequences that span an internal loop, a connecting stem, and a terminal hairpin loop (Fig. 4). The conserved sequences occur in only one permutation, suggesting that the sequences participate in a single complex overall structure. The broad and obvious conservation of these sequences in many independent cases is *prima facie* evidence for their function in the activity selected, free L-histidine binding. However, the role of each of the conserved sequences is supported independently and in detail by chemical accessibility/interference data on 13 independent versions of the site (Fig. 6). In addition, mutational replacements nucleotides in the 5′, the middle, or the 3′ conserved regions of one sequence (in a minimal fragment of isolate 945; Fig. 7) all inactivate histidine binding. The sequence requirements for affinity are accordingly relatively well defined.

Furthermore, this is probably the simplest RNA solution to protonated histidine affinity under selective conditions, because the recovery of the same site in so many independent ways, as 83% of the pooled

sequences (and without the appearance of a single alternative active structure) suggests that this site's dominance requires a clear rationale. That it is the simplest site would be such a rationale; the frequency of a site rises exponentially as the number of required nucleotides becomes smaller (Knight and Yarus 2003; Yarus and Knight 2004).

This is particularly remarkable because this dominant site for L-histidine contains among its conserved sequences two anticodons for the cognate amino acid, G/AUG (codons CAY; see Fig. 6). The flanking nucleotides of the 5′ GUG (internal loop) are completely conserved, while the middle U occurs in 62% of families (Fig. 6B). All three nucleotides of GUG give a chemical signature characteristic of active binding site nucleotides. Further, replacement of two, or all, of these nucleotides completely inactivates histidine binding in a minimal site (Fig. 7). The 3′ AUG (hairpin loop) anticodon nucleotides are 71–85–100% conserved, and the flanking nucleotides give protection/interference signatures. The middle U is chemically silent, and the uracil base may therefore not participate directly in binding. However, mutation of two of three, or all, of these AUG nucleotides inactivates the site (Fig. 7). Thus conservation, chemical probing/interference, and mutational data agree that the anticodon nucleotides are implicated in histidine binding.

Accordingly, just as for isoleucine (Lozupone et al. 2003; Majerfeld and Yarus 1998) and for tryptophan (Yarus et al. 2005), the simplest and most abundant site for an amino acid contains conserved coding triplets. Therefore, selection for isoleucine, histidine, or tryptophan affinity in RNA-like molecules would have been sufficient to associate sequences that became coding triplets with the amino acid itself.

Acknowledgments. We thank members or our laboratory for comments on the manuscript. Preparation of this manuscript was supported by NIH Grant GM 48080 and NASA Center for Astrobiology Grant NCC2-1052.

References

Burgstaller P, Kochoyan M, Famulok M (1995) Structural probing and damage selection of citrulline- and arginine-specific RNA aptamers identify base positions required for binding. Nucleic Acids Research 23:4769–4776

Cadwell RC, Joyce GF (1994) Mutagenic PCR. PCR Meth Appl 2:S136–S140

Ciesiolka J, Illangasekare M, Majerfeld I, Nickles T, Welch M, Yarus M, Zinnen S (1996) Affinity selection-amplification from randomized ribooligonucleotide pools. Methods Enzymol 267:315–335

Connell GJ, Yarus M (1994) RNAs with dual specificity and dual RNAs with similar specificity. Science 264:1137–1141

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

Famulok M (1994) Molecular recgnition of amino acids by RNA-aptamers: An L-citrulline binding RNA motif and its evolution into an L-arginine binder. J Am Chem Soc 116:1698–1706

Famulok M, Szostak JW (1992) Stereospecific recognition of tryptophan agarose by in vitro selected RNA. J Am Chem Soc 114:3990–3991

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

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

Knight R, Yarus M (2003) Finding specific RNA motifs: Function in a zeptomole world? RNA 9:218–230

Knight R, Birmingham A, Yarus M (2004) BayesFold: Rational second-degree folds that combine thermodynamic, covariation, and chemical data for aligned RNA sequences. RNA 10:1323–1336

Knight RD, Landweber LF (1998) Rhyme or reason: RNA-arginine interactions and the genetic code. Chem Biol 5 :R215–220

Krol A, Carbon P (1989) A guide for probing native small nuclear RNA and ribonucleoprotein structure. Meth Enzymol 180:212–227

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

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 associated coding sequences. RNA 4:471–478

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

Morris KN, Jensen KB, Julin CM, Weil M, Gold L (1998) High affinity ligands from in vitro selection: Complex targets. Proc Natl Acad Sci USA 95:2902–2907

Roth A, Breaker RR (1998) An amino acid as a cofactor for a catalytic polynucleotide. Proc Natl Acad Sci USA 95:6027–6031

Wilson DS, Keefe AD (2000) Random metagenesis by PCR. In: Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K (eds) Current protocols in molecular biology. John Wiley and Sons, New York, pp 831–839

Woese CR, Dugre DH, Saxinger WC, Dugre SA (1966) The molecular basis for the genetic code. Proc Natl Acad Sci USA 55:966–974

Yang Y, Kochoyan M, Burgstaller P, Westhof E, Famulok M (1996) Structural basis of ligand discrimination by two related RNA aptamers resolved by NMR spectroscopy. Science 272:1343–1347

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

Yarus M, Christian EL (1989) Genetic code origins. Nature 342:349–350

Yarus M, Knight RD (2004) The scope of selection. In: Pouplana LR (ed) The genetic code and the origin of life. Landes Bioscence, Georgetown, TX, pp 75–91

Yarus M, Caporaso JG, Knight R (February 11, 2005) Origins of the genetic code: The escaped triplet theory. Annu Res Biochem 74:179–198. DOI: 10.1146/annurev.biochem.74.082803.13119