A diminutive and specific RNA binding site for L-tryptophan

Irene Majerfeld and Michael Yarus*

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

Received July 20, 2005; Revised and Accepted September 7, 2005

ABSTRACT

Selection for amino acid affinity by elution of RNAs from tryptophan–Sepharose using free L-tryptophan evokes one sequence predominantly (\(K_D = 12 \mu\text{M}\)), a symmetrical internal loop of 3 nt per side. Though we have also isolated larger sequences with affinity for tryptophan, successively squeezed selection in randomized tracts of 70, 60, 40, 20 and 17 nt show that this internal loop is the simplest sequence that can meet the column affinity selection. From sequence variation in ~50 independent isolates, only 26 bits of information are required to describe this loop (equivalent to only 13 fully conserved nucleotides). Thus, it is among the simplest amino acid binding sites known, as well as selective among hydrophobic side chains. Among site sequences defined as essential to affinity by conservation, protection and modification-interference, there is a recurring CCA sequence (a tryptophan anticodon triplet) which apparently forms one side of the binding site. Such conserved juxtaposition of tryptophan with a cognate coding triplet supports a stereochemical origin for the genetic code.

INTRODUCTION

Stereochemical theories for the origin of the genetic code propose chemical affinity between amino acid and RNA as the basis for association between codons and/or anticodons and cognate amino acids (1). Recently, Yarus et al. (2) have reviewed the related hypothesis that code assignments originated as part of RNA or RNA-like amino acid binding sites and triplets subsequently ‘escaped’ (took on new functions) to participate as codons and anticodons in a more modern translation apparatus. By purifying a minority of functional molecules from \(10^{14}\) to \(10^{15}\) mostly inactive random sequences, selection-amplification or SELEX (3) can be used to experimentally explore possible RNA–amino acid associations, perhaps repeating nature’s ancient experiment with the genetic code. RNA binding sites for 8 of the standard 20 encoded amino acids have presently been isolated and characterized. Here, we describe the simplest stereoselective and side chain specific amino acid binding site we have seen, for L-Trp. Notably, this site fits the overall stereochemical pattern (2).

MATERIALS AND METHODS

Selection procedures

The coupling of amino acids to EAH Sepharose 4B (Amersham) for the preparation of a mixed affinity matrix for the amino acids glutamine, histidine, tryptophan and valine (0.4 mM each) has been described previously (4). Two multitarget selections, I and II, using a 70 nt randomized region have also been described previously (4). Two multitarget selections, I and II, using a 70 nt randomized region have also been described previously (4). Selections from 60, 40, 20 and 17 nt randomized regions were performed on 0.4 mM tryptophan–Sepharose following the protocol described for selection II (4) except: (i) they did not include mutagenic PCR and (ii) a proofreading polymerase, Pfu Turbo (Stratagene), was used for amplification. In all selections, except for the first round, selections were preceded by a counterselection on acetylated Sepharose. The initial RNA pool for the 60 and 40N selections were transcribed from \(24\cdot10^{14}\) DNA sequences; selections from 20 and 17N random lengths were started from \(24\cdot10^{14}\) DNA sequences. Selection buffer was 50 mM HEPES (pH 7.0), 250 mM NaCl, 5 mM each MgCl\(_2\) and CaCl\(_2\) and glycine at a concentration equal to that of the eluting amino acid. The same buffer, without glycine, was used in other experiments.

The initial DNA sequences (T7 promoter underlined) were as follows. Selection from 70 randomized positions, taatac-gac-ctatagaggtcacttaagctctatcgg N(70) aaagcggcctagcgatcga. Selection from 60 and 40 randomized positions, taatac-gac-ctatagaggtcacttaagctctatcgg N(60) aaagcggcctagcgatcga. Selection from 20 randomized positions, taatac-gac-ctatagaggtcacttaagctctatcgg N(20) aaagcggcctagcgatcga.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oxfordjournals.org
Selection from 17 randomized positions, taatacgactcataagggactgaagtt N(17)gcatgagtagtctgacagca.

Doped selection
For the initial pool, the 34 nt long minimal functional structure of isolate Trp 20–605, nucleotides A10–U43, was mutagenized to 30% per position and new primers were added. The selection protocol was the same as for the original selection. The initial DNA sequence was taatacgactcataagggactgaagttgaagtgacatca (34 doped) cgatgtcggagaggagatgcg.

Boundary determination
5' and 3' end-labeled RNA was partially hydrolyzed with 50 mM NaOH for 30 s at 85°C. After precipitation, folded RNA was fractionated through the affinity column and run-through and eluted fractions were analyzed by PAGE.

K_D determinations
Dissociation constants (K_D) were determined by affinity chromatography, ultrafiltration and equilibrium dialysis. Affinity chromatography K_D is obtained by isocratic elution from the affinity matrix with and without ligand using K_D = 1 (V_el − V_e)/V_e − V_e) (5), where V_el and V_e are the median elution volumes of the RNA in the presence and absence of ligand and V_n is the median elution volume of RNA in the absence of any affinity.

In ultrafiltration K_D determinations (6), ~15 µl from a 100 µl tritiated ligand-RNA solution were forced by centrifugation through a Microcon YM30 (Amicon) filter, thereby fractionating free from (free plus bound) ligand.

In equilibrium dialysis 50 µl RNA and ligand samples were equilibrated overnight across the dialysis membrane of a Nist Group SDIS 100KE disposable dialyzer (Southborough, MA).

RESULTS
Selection
Multitarget column affinity selections for RNA binding sites directed at the amino acids glutamine, histidine, tryptophan and valine produced aptamers for histidine and tryptophan, but exhausted the sequence variety of the pool without evidence of specific affinity for the other two amino acids. Experimental details of these selections have been described, along with the simplest L-histidine-binding RNA (4). We now present the tryptophan sites and show that the simplest possible is a conserved symmetrical internal loop, 3 over 3 nt.

In an initial pool of ~10^14 unique sequences with a 70 nt long randomized region, histidine-specific aptamers were isolated from selection cycles 6 and 7 (4). Tryptophan-specific aptamers were obtained from selection cycle 8. At this round of selection, 18 and 17% of the RNA applied to the column was eluted by 1 or 0.2 mM tryptophan (selections I and II, respectively).

Figure 1 is a summary of the selected sequences; the initially randomized section of the sequences is shown and

![Figure 1. Sequence composition of the selected pool. The variable region of one typical representative from each family and the number of isolates are shown. (A) Sequences that contain the CYA Trp motif. Alignment is based on module 1. Shaded nucleotides are at least 81% conserved. (B) Sequences that do not contain the CYA motif. A superscript asterisk following the sequence label indicates that dissociation constants (K_D) were determined for that clone (Table 1). Nineteen unique sequences that do not contain the CYA motif are not shown.](image-url)
families (sequences apparently enriched from one parental RNA) are represented by one typical member of the group. The number of sequences in each family is indicated at the right. A total of 158 clones were sequenced. Eighty-one percent of the pool or 128 sequences, shown in Figure 1A, share two conserved sequences, shaded in Figure 1A and called module 1 and module 2. Module 1, on which sequences are aligned in Figure 1A, is 5'-RRGACCG, and module 2 is the sequence 5'-CGCYACY. Module 2 occurs in both permutations, 5' or 3' to module 1. Common folds (7) shared by different sequences containing this motif place nucleotides GAC (module 1) and nucleotides CYA (module 2) in a small internal loop framed by two partially conserved stems. We will show below that these very conserved segments (shaded nucleotides in Figure 1 are at least 81% conserved) form the predominant and simplest tryptophan binding site that we will call the CYA motif.

The pool contains 4 sequences, shown in Figure 1B, that were enriched into families of 2–4 sequences that do not contain the above described conserved segments. However, they fulfill the selection requirement of binding tryptophan. Nineteen unique sequences (12% of the pool) that do not contain the conserved elements were not tested and are not presented in Figure 1.

In summary, more than one RNA folding fulfilled the selection requirement. There seem to be several ways to construct an RNA site with affinity for tryptophan. However, one motif, containing a small internal loop, predominates in the 70 nt selected pool. It was isolated independently 19 times and comprised 81% of RNAs in the population.

**Binding by selected RNAs**

Figure 2A and B illustrates the binding properties and specificity of the various families and single sequences that carry the predominant motif shown in Figure 1A. Trp 70–317 has a CUA counterloop, Trp 70–727 a CCA counterloop. Their behavior is very similar, binding well and eluting quantitatively from the tryptophan-Sepharose column. The binding is specific since other large hydrophobic amino acids with aliphatic or aromatic side chains at a concentration about 100-fold over $K_D$ for tryptophan (see below) do not elute bound RNA. The addition of tryptophan to the column buffer quickly and quantitatively removes the bound RNA in Figure 2, leaving none to be recovered by denaturing the RNA in subsequent Na$^+$ and EDTA buffer.

The same properties of good binding and specificity for the amino acid side chain are characteristics of Trp 70–585 (Figure 2C), an aptamer with an alternative binding site.

Dissociation constants ($K_D$) for L- and D-tryptophan were determined by isocratic elution from tryptophan-Sepharose (Table 1). This procedure determines $K_D$ for the ligand in solution, the influence of the matrix on the measurement being normalized by a reference measurement of the median RNA elution volume on the same column in the absence of free ligand (5). The procedure is equivalent to equilibrium dialysis in that the measured difference between bound (fixed) and unbound (movable) species is generated by an interposed solid phase. For five isolates, dissociation constants for L-tryptophan were also determined by ultrafiltration, a quasi-equilibrium method. Finally, for Trp 70–727 we used equilibrium dialysis as well. Values obtained by these three alternative methods (Table 1) show no systematic discrepancy; therefore, all apparently determine the same $K_D$. However, affinity chromatography is uniquely useful because it allows convenient $K_D$ measurement for the D-amino acid and other congeners not available in radioactive form.

$K_D$ for L-tryptophan ranged between 7 and 25 $\mu$M and for D-tryptophan between 1 and 3 mM. The discrimination
between L- and D- amino acid therefore ranged between 67- and 280-fold, though stereoselectivity was not selected during isolation of these RNAs. As might be expected (since they were obtained via the same selection regimen), these $K_D$ are similar to those of the previously described histidine aptamer (4).

For additional insight into the recognition of the amino acid ligand by the CYA motif, we determined $K_D$ for some analogs of tryptophan using two isolates of the major motif, Trp70–317 and Trp 70–188, the second a member of the Trp 70–305 family (Figure 1). The ester of tryptophan (H-Trp-O-Me) and 4-methyl-DL-Trp, with a substitution on the indole ring, are effective ligands, with $K_D$ increased by a factor of 2 or less ($\Delta \Delta G \approx 0.4$ kcal/mol). The removal of the carboxyl group was of more consequence since $K_D$ for tryptamine was 4–7 times higher than that of tryptophan ($\Delta \Delta G = 0.8–1.2$ kcal/mol). The most substantial effect was caused by substitution at the $\alpha$-carbon, $\alpha$-Methyl-DL-Trp had a $K_D$ at least 60-fold higher than that of tryptophan ($\Delta \Delta G = 2.4$ kcal/mol or higher), assuming that only the L-congener is active (compare Table 1). Tryptophan binding by the major loop motif evidently depends on close contact with both the side chain and the $\alpha$-carbon and its substituents.

### Identification of the binding site nucleotides

The apparent conservation in the pool restricted to a small internal loop and short supporting stems with extensive variability beyond that suggests a small, simple binding site. To confirm the size and location of the binding site, we chemically probed several isolates from the pool. We used chemical modification with DMS (for A and C) and CMCT (for G and U) followed by reverse transcription (8) to identify nucleotides with reactivities affected by tryptophan binding. Ten different CYA motif isolates with both CUA and CCA counterloops were tested. Two isolates, Trp70–585 and Trp 70–358, that do not contain the predominant motif were also tested. In addition, modification-interference was compared for two isolates containing the alternative counterloops, Trp 70–93 (CUA) and Trp 70–727 (CCA).

Figure 3A and B shows the protections and enhancements obtained for Trp 70–305 (CUA counterloop) and Trp 70–727 (CCA counterloop). Data are graphically summarized in Figure 4A and B over BayesFold (7) predictions for the most likely secondary structure in view of the allowed variation in aligned sequences and the DMS and CMCT chemical
accessibility data. Nucleotides with reactivities affected by bound tryptophan concentrate in the conserved loop, with strong protections on one side and a strong enhancement at the conserved A of the counterloop. Interferences for Trp 70–727 confirm the essential function of the conserved nucleotides and extend the site to other nucleotides in the conserved loop, as well as implying a role for a few others outside the loop (Figure 4B).
Figure 5 summarizes the data for all clones tested. The loop nucleotides sensitive to ligand binding approximately reproduce in all clones tested, whether they have a CUA or a CCA counterloop, emphasizing the similar nucleotide function in independently isolated sequences and reinforcing the relevance of the conservation data shown in Figure 1. There seems no doubt that we observed the recurrence of the same site and binding mechanism, though in two permutations and with different spacing between conserved elements.

As a complementary approach to defining the binding site we determined the 3′ and 5′ boundaries for minimal binding structure in Trp 70–727 by partial alkaline hydrolysis. These boundaries, which include the conserved loop, are indicated in Figure 4B. A fragment was constructed observing these limits, inverting the C26 G69 base pair to facilitate transcription. The resulting 44 nt fragment had the binding properties of the parental isolate with a $K_D$ for tryptophan of 15 $\mu$M, indistinguishable from $K_D = 13$ $\mu$M measured for the full-length parental Trp 70–727 sequence (Table 1).

Figure 4 also shows chemical data for two isolates that do not contain the most frequent tryptophan site. Nucleotides changed by amino acid binding are located in a symmetrical loop in Trp 70–730, a two sequence family (Figure 4C). This 8 nt loop appears to be a variation of the major motif loop, in which the sequence AAC substitutes for the conserved GAC of module 1. Similar protections and enhancement support the relationship. Affinity chromatography for Trp 70–730 was shown in Figure 2C. For Trp 70–358, DMS and CMCT probing data suggest an 11 nt hairpin loop as the location for bound tryptophan (Figure 4D).

Selections from random regions of different lengths

Nineteen independent isolations of the same CYA motif, as well as its majority among selected RNAs suggest that it is the simplest structure that can bind free L-tryptophan and meet a column affinity selection. Simplicity and its implied result, dominance of any affinity selection, are pertinent to theories of the origin of the genetic code. Lozupone et al. (9) isolated the simplest isoleucine aptamer by selecting for the activity from progressively shorter random segments until no activity could be found. The simplest motif, requiring the least number of nucleotides, should persist as the randomized tract shortens and selection for small size intensifies; it should be the last one to disappear. We followed this ‘successively squeezed selection’ protocol and repeated affinity selection on 0.4 mM tryptophan–Sepharose starting from RNAs with 60, 40, 20 and 17 consecutive randomized nucleotides. The selections at 20 and 17 nt used 500 pmol of random sequence DNA as template for the transcription of initial RNA, so these selections probably had full sequence representation (contained every possible sequence) for the starting 17 and 20mers. Thus, if any alternative 17mer or 20mer ribonucleotide can bind tryptophan, we expect to have tested it.

Table 2 summarizes these four new selections and also includes comparable data from the original 70 randomized sequences with other binding sites.
nucleotide selection. The CYA motif was predominant in the selections from 60, 40 and 20 random nucleotides, constituting 67, 72 and 63% of the sequences in the final pools while the internal loop was not isolated from the 17 nt selection. Figure 6 summarizes the sequences from all selections using shortened random regions. Figure 6A lists the sequences that contain the CYA motif. Because a proofreading DNA polymerase was used in the amplification step of these selections, single nucleotide changes are treated as possible independent isolations and their full sequences are shown. However, because errors could also have occurred during reverse transcription, the origin of sequences differing in only few nucleotides is uncertain. In particular, sequence Trp 40–164 differs only in two positions from sequence Trp 40–102. To treat this uncertainty in a balanced way, sequences which differ in 3 or fewer nucleotides are not classified as of certain independent origin and a range is shown in Table 2.

We tested 13 sequences from the three successful selections for binding and elution from the affinity column and/or for specificity in comparison with other hydrophobic amino acids and for D-tryptophan; they behave, as expected, similarly to isolates from the original 70 nt selection. In every independent isolation of the CYA motif from the 20 random selection the binding loop was generated from initially randomized nucleotides. However, this length is not enough to generate a stable structure, and nucleotides from the primers were used to complete a stem (Figure 7A).

Other selected sequences

Here, we discuss sequences selected from the 60, 40 and 17 random nucleotide selections other than the predominant CYA motif. Figure 6B lists these sequences that have been enriched into multi-sequence families or appear in groups of related sequences with more than one possible origin; this latter class is noted by an asterisk. Unique sequences are not included.

Two minority sequences, Trp 60–216 from the 60 nt selection and Trp 40–126 from the 40 nt selection, representing 12.5 and 2.5% of their respective pools share a 10 nt long segment, underlined in Figure 6B. Trp 40–126 shows binding and specificity comparable with that of sequences with the CYA motif (data not shown). BayesFold secondary structure predictions for these two sequences place the underlined conserved nucleotides AAAAC in an internal loop for which the counterloop, CUAAG in both sequences, is mostly drawn from the 3’ fixed region (see Materials and Methods for primer sequence). This loop is larger but shows a resemblance to the internal loop of sequence Trp 70–585, which chemical probing suggests as the site of binding (Figures 4 and 5). Perhaps these minority sequences are related, but contain an apparently larger internal loop site than that of the predominant CYA motif.

Three groups of sequences in the 20 nt selection do not contain the CYA motif (Figure 6B); they were tested by affinity chromatography for tryptophan binding. Clone Trp 20–605 represents a group of five sequences (10% of the pool) with good binding and specificity as illustrated in Figure 8A. L-tryptophan effectively elutes RNA bound to the column that L-tyrosine or D-tryptophan cannot. Neither phenylalanine nor isoleucine elutes bound RNA (data not shown). Trp 20–605, for which a predicted fold is shown in Figure 7B, represents an alternative binding site to be discussed below. Trp 20–604 and Trp 20–623 do not bind as well to the column and have minimal or no affinity for free tryptophan (Figure 8B and C). Their column behavior as well as their persistence in the pool is probably explained by affinity for the matrix itself.

The selection from 17 random nucleotides did not result in a tryptophan elution peak in nine cycles of selection (Table 2). However, this pool was cloned and sequenced. Twenty-six sequences showed minimal variability, indicating that sequence variation had been depleted by selection. Two major groups of sequences dominated the pool (Figure 6B). A family represented by Trp 17–417 was present in 17 copies (65% of the pool), and a group of six related sequences with 2–4 possibly different origins made up 23% of the pool. There were also three unique sequences. The major sequences were tested by chromatography. Trp 17–417 shows only residual binding to either the acetylated Sepharose used for counterselection (see section 2.1) or to tryptophan–Sepharose (Figure 8D). Its enrichment may be due to an advantage in amplification or transcription. Trp 17–435 binds tightly to both acetylated and tryptophan–Sepharose (Figure 8E) but shows no specificity for the selection ligands and is instead removed from the column by a high salt wash. Thus, there were no true aptamers in the 17 nt selection.

In summary, in every selection that yielded tryptophan aptamers the CYA motif was the most abundant result, comprising 63–81% of the total RNA. Alternative binding sites never surpassed 12%, suggesting they were less abundant in the initial pool and therefore possibly more complex. In the selection from 20 randomized nucleotides, the CYA motif appeared even though it could not be completely constructed within the initially randomized nucleotides. A further reduction by three randomized positions removed all possibility of forming either the CYA motif or any other functional site. In addition to the majority CYA motif, the selection from 20 randomized nucleotides produced a new tryptophan aptamer, Trp 20–605, comparable in affinity and specificity to the CYA site. We analyze this new site in the following section.

The alternative site in aptamer Trp 20–605: characterization and doped selection

Trp 20–605 represents a group of five sequences (10% of the corresponding pool) with 1–3 apparently independent origins (Figure 6B, 7B). No significant similarity was detected by PILEUP (Wisconsin Package Version 10.1) between Trp 20–605 and any other isolate from any other selection. Our purpose requires that we determine the size of this new active site to see if it is simpler than the CYA motif. Chemical probing with DMS and CMCT identified several positions.

| Table 2. Summary of selections |
|-----------------------------|
| Initially randomized length | 70  60  40  20  17 |
| Final pool at selection cycle | 8   4   4   5   9 |
| RNA recovery in final elution | 18/17% | 26% | 34% | 17% | 5% |
| Number of clones sequenced | 158 | 48 | 50 | 49 | 26 |
| Abundance of the CYA motif | 81% | 67% | 72% | 63% | 0% |
| Independent isolations, CYA motif | 19 | 11 | 15–16b | 4–6b | 0 |

aThe two figures correspond to two parallel selections as described in Materials and Methods.
bSee text for explanation of range of possible independent isolations.
with ligand-dependent sensitivities. They localize in a terminal loop and extend into an adjacent asymmetrical loop that includes nucleotides from the selection primers (Figure 7B). We also determined the minimal active molecule by limited alkaline hydrolysis, the resulting 3' and 5' boundaries on nucleotides 10 and 43, respectively, are shown in Figure 7B. In confirmation of the chemical data, conservations encompass the nucleotides highlighted by chemical probing. Based on this result, a 34 nt fragment (with minimal nucleotide changes for efficient T7 RNA polymerase transcription) was tested by

| Isolate Number | Randomized part of sequence | No. of independent origins |
|----------------|-----------------------------|-----------------------------|
| **A. Sequences containing the CYA motif** |
| From 60 random |                             |                             |
| Trp 60-203    | UAGCAGCUUGCAGCUAGCAGCUAG    | 14                          |
| Trp 60-204    | CAAGCUAGCUAGCUAGCUAGCUAG    | 7                           |
| Trp 60-218b   | CAAGCUAGCUAGCUAGCUAGCUAG    | 2                           |
| Trp 60-207    | CAAGCUAGCUAGCUAGCUAGCUAG    | 2                           |
| Trp 60-236    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 60-203    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 60-209b   | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 60-231    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 60-229    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 60-202    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 60-285    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| From 40 random |                             |                             |
| Trp 40-102    | CAAGCUAGCUAGCUAGCUAGCUAG    | 17                          |
| Trp 40-122    | CAAGCUAGCUAGCUAGCUAGCUAG    | 2                           |
| Trp 40-164    | CAAGCUAGCUAGCUAGCUAGCUAG    | 2                           |
| Trp 40-156    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 40-120    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 40-11    | AGACUAGCUAGCUAGCUAGCUAG     | 1                           |
| Trp 40-145    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 40-137    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 40-143    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 40-129    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 40-123    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 40-105    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 40-166    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| From 20 random |                             |                             |
| Trp 20-636    | CAAGCUAGCUAGCUAGCUAGCUAG    | 21                          |
| Trp 20-642b   | CAAGCUAGCUAGCUAGCUAGCUAG    | 2                           |
| Trp 20-616c   | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 20-637    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 20-613    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| Trp 20-607    | CAAGCUAGCUAGCUAGCUAGCUAG    | 1                           |
| **B. Enriched sequences and groups of related sequences with more than one origin that do not contain the motif** |
| From 60 random |                             |                             |
| Trp 60-216    | AGAAUGAAUAGUAGUAGUAGUAGUAG | 6                           |
| Trp 60-205    | AGAAUGAAUAGUAGUAGUAGUAGUAG | 2                           |
| From 40 random |                             |                             |
| Trp 40-126    | AGAAUGAAUAGUAGUAGUAGUAGUAG | 3                           |
| Trp 40-144    | AGAAUGAAUAGUAGUAGUAGUAGUAG | 2                           |
| From 20 random |                             |                             |
| Trp 20-605*   | AGAAUGAAUAGUAGUAGUAGUAGUAG | 3                           |
| Trp 20-604*   | AGAAUGAAUAGUAGUAGUAGUAGUAG | 5                           |
| Trp 20-623*   | AGAAUGAAUAGUAGUAGUAGUAGUAG | 4                           |
| From 17 random |                             |                             |
| Trp 17-417    | AGAAUGAAUAGUAGUAGUAGUAGUAG | 17                          |
| Trp 17-435*   | AGAAUGAAUAGUAGUAGUAGUAGUAG | 6                           |

Figure 6. Summary of sequences from the 60, 40, 20 and 17 random selections. The randomized regions of a typical representative from each family and the number of isolates are shown. (A) Sequences that contain the CYA motif. Sequences are aligned on module 1. Shading indicates conserved nucleotides (minimum 81%), same as in Figure 1. Because a proofreading polymerase was used for the amplification step of the selections, single nucleotide differences are treated as independent isolations and are shown independently, but see text. (B) Sequences that do not contain the CYA motif. Asterisk indicates a group of sequences with one or few nucleotide variations that for the sake of brevity are not shown individually. The underlined segments of Trp 60–216 and Trp 40–126 are the same.
affinity chromatography and found to have affinity similar to the parent RNA.

To identify all the positions essential for binding, the 34 nt active fragment was mutagenized to 30% (nucleotides 10–43 in Figure 7B). With the addition of new primers, the active structure was re-selected using the same procedure as before. From a starting RNA population transcribed from 2 × 10^{14} independent DNA sequences, an elution peak of 20% was obtained after four cycles of selection. This tryptophan-eluted pool was cloned and 59 sequences were obtained.

The nucleotide variation and consensus sequence within a 26 nt long segment is shown in Figure 9B where shaded positions are at least 70% conserved. Frequencies are adjusted for the initial bias (7:1:1:1) in the doped pool. Figure 9B also shows the relevant segment of a BayesFold prediction for the structure common to the 59 available aligned sequences. The re-mutagenized and re-selected RNAs recapitulated the parental loop structure of Trp 20–605 (Figure 7), as we had intended.

There are seven invariant positions and 16 nt at least 70% conserved. The highest conservation is in the terminal loop, nucleotides 36–42, where five out of seven positions are invariant and one, position 37, is always a purine with a strong preference for an A. The very high conservation extends to the base pair closing the loop. The involvement of the conserved loop sequence in binding, specially the terminal loop, has already been demonstrated by chemical probing of Trp 20–605 (Figure 7).

We compared the complexity of the Trp 20–605 site with that of the predominant CYA motif by calculating the information content required to specify each structure. The observed Shannon uncertainty ($H_{\text{obs}}$) for loop and stem positions was calculated as described by Legiewicz and Yarus (10) according to the method of Schneider et al. (11). $H_{\text{obs}}$ was subtracted from the maximum Shannon uncertainty ($H_{\text{max}}$) to obtain the information content at that position (bits per position). The total information content was obtained by adding the information content for all positions. We used the nucleotide variation data and the structures shown in Figure 9A and B for these calculations.

The Trp 20–605 hairpin is much more complex, requiring 39 bits versus 26 bits required to specify the predominant CYA motif. Thus, the CYA motif is clearly the simplest tryptophan aptamer by all criteria.

The Trp 20–605 motif was isolated three times and its sequences represent 10% of the selected 20 nt pool. The CYA motif was isolated seven times and constitutes 63% of the 20 nt pool. The large difference in information content between the two can easily explain the predominance of the CYA motif. In addition to reduced size, the modularity of the CYA motif would increase its frequency in the initial pool (12). The abundance of the Trp 20–605 motif in the pool is likely due to the accidental match between the constant sequences we chose and the sequence requirements of the site (compare Figures 7B and 9B). This reduced the number of randomized nucleotides which were needed to construct the Trp 20–605 motif and made it more probable. Conceivably, another contributing factor may be the shortness of the initially randomized region that forced the use of suboptimal nucleotides in the construction of the smaller motif. This sequence of events is similar to that of Legiewicz and Yarus (10) for isolation of a larger, less probable isoleucine site because of the contribution from accidentally incorporated constant sequences.

Figure 7. BayesFold secondary structure predictions for isolates from the 20 nt selection. The length of the initially random region is indicated. (A) Trp 20–625, a CYA motif clone. (B) Trp 20–605. The calculation for this folding used chemical accessibility data. DMS and CMCT chemical probing data are shown; symbols are as in Figure 4. 5′ and 3′ boundaries for a functional structure determined by partial alkaline hydrolysis are indicated.
DISCUSSION

We believe that we have isolated the simplest RNA structure that can bind L-tryptophan by an affinity chromatographic criterion. This conclusion rests on its recurrent abundance in independent selections, and more particularly, on its majority when the randomized tract is shortened until the selection fails. Figure 9A shows the consensus sequence and nucleotide conservation for the CYA motif, as calculated from four independent successful selections and 228 sequences. Module 1 and module 2 have the consensus sequences 5'-RRGACCG and 5'-CGCYACY, respectively, with all nucleotides at least 81% conserved. The nucleotides forming the small symmetrical loop, GAC on the one side and CYA on the counterloop, are at least 99.6% conserved. The middle nucleotide of the counterloop can be C and U yielding functionally indistinguishable aptamers. Nucleotides closing the loop are also highly conserved. Chemical probing implicated the internal loop as the site of binding, with protections and/or interferences in most of the loop positions.

The structure of the CYA motif, as well as its predominance among selected sequences is observed in these experiments despite the use of several different sets of primers/constant flanking sequences. Thus, its identification as the simplest active sequence likely does not depend on a special interaction with constant sequences (9).

Small functional internal loops are common within in vitro selected amino acid aptamers, and form the binding pocket in several arginine (13–17), and one isoleucine aptamer (9,18). This particular tryptophan binding CYA loop has been mentioned previously. The minimal functional fragment of RNA Trp 70–727 described above was used as the amino acid specific element in the construction of a side chain specific RNA amino acid transporter (19), capable of speeding the equilibration of L-tryptophan across the bilayer membrane of phospholipid vesicles.

Other RNAs with a relation to tryptophan have been also reported, often with concurrent affinity for other aromatic amino acids. A stereospecific aptamer for D-tryptophan–Sepharose was isolated by affinity chromatography from a 120 randomized nucleotide region (20), though free amino acid was apparently not an effective ligand. In another selection for L-phenylalanine affinity, the major selected sites accepted tryptophan in addition to the selection target, though with no stereospecificity (21). Another L-phenylalanine affinity selection resulted in phenylalanine-specific aptamers as well as a class with general L-aromatic amino acid specificity (22). A mutagenized sequence with affinity for dopamine, when re-selected for tyrosine, yielded three independent L-tyrosine sites that also bound L-tryptophan. Two of these accepted multiple aromatic amino acids, though one of the three moderately discriminated phenylalanine (23). Accordingly, RNA can easily specify recognition of aromatic side chains, doing so using differing structures. In light of this
The tryptophan anticodon, CCA, appears in one of the two counterloops. It occurs in 20% of total sequences (Figure 9A) and in 14–15 out of 49–53 possible independent isolations of the CYA motif (Table 1). Therefore, just as for L-isoleucine (9,18) and L-histidine (4), selection of tryptophan affinity produces a cognate coding triplet, CCA, as a conserved part of the most easily recurring amino acid binding site. The small size of the active motif suggests in this case that the CCA would be in close proximity to a bound amino acid. The alternative counterloop has a CUA triplet, complementary to UAG, an anticodon triplet which did not enter the standard code as an amino acid. Accordingly, these data support a stereochemical origin for the genetic code for tryptophan, as posited by the escaped triplet theory (2).

ACKNOWLEDGEMENTS

The authors thank members of the Yarus laboratory for comments on the manuscript. This work was supported by NIH RG GM48080 and by NASA Astrobiology Center NCC2-1052. Funding to pay the Open Access publication charges for this article was provided by the same sources.

Conflict of interest statement. None declared.

REFERENCES

1. Woese,C.R. (1965) On the evolution of the genetic code. Proc. Natl Acad. Sci. USA, 54, 1546–1552.
2. Yarus,M., Caporaso,J.G. and Knight,R. (2005) The escaped triplet theory of the genetic code. Annu. Rev. Biochem., 74, 179–198.
3. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science, 249, 505–510.
4. Majerfeld,I., Puthenvedu,D. and Yarus,M. (2005) RNA affinity for molecular L-histidine; genetic code origins. J. Mol. Evol., 61, 226–235.
5. Ciesiolka,J., Illangasekare,M., Majerfeld,I., Nickles,T., Welch,M., Yarus,M. and Zinnen,S. (1996) Affinity selection-amplification from randomized ribo oligonucleotide pools. Methods Enzymol., 267, 315–335.
6. Dean,A.M., Lee,M.H. and Koshland,D.E.,Jr (1989) Phosphorylation inactivates Escherichia coli isocitrate dehydrogenase by preventing isocitrate binding. J. Biol. Chem., 264, 20482–20486.
7. Knight,R., Birmingham,A. and Yarus,M. (2004) BayesFold: rational secondary folds that combine thermodynamic, covariation, and chemical data for aligned RNA sequences. RNA, 10, 1323–1336.
8. Krolikowski, A. and Carbon, P. (1989) A guide for probing native small nuclear RNA and ribonucleoprotein structure. *Methods Enzymol.*, 180, 212–227.

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

10. Legiewicz, M. and Yarus, M. (2005) A more complex isoleucine aptamer with a cognate triplet. *J. Biol. Chem.*, 280, 19815–19822.

11. Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188, 415–431.

12. Knight, R., and Yarus, M. (2003) Finding specific RNA motifs: function in a zeptomole world? *RNA*, 9, 218–230.

13. Connell, G. J., Illangasekare, M., and Yarus, M. (1993) Three small ribooligonucleotides with specific arginine sites. *Biochemistry*, 32, 5497–5502.

14. Connell, G. J., and Yarus, M. (1994) RNAs with dual specificity and dual RNAs with similar specificity. *Science*, 264, 1137–1141.

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

16. Tao, J., and Frankel, A. D. (1996) Arginine-binding RNAs resembling TAR identified by *in vitro* selection. *Biochemistry*, 35, 2229–2238.

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

18. Majerfeld, I., and Yarus, M. (1998) Isoleucine:RNA sites with associated coding sequences. *RNA*, 4, 471–478.

19. Janas, T., Janas, T., and Yarus, M. (2004) A membrane transporter for tryptophan composed of RNA. *RNA*, 10, 1541–1549.

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

21. Zinnen, S., and Yarus, M. (1995) An RNA pocket for the planar aromatic side chains of phenylalanine and tryptophane. *Nucleic Acids Symp Ser.*, 148–151.

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

23. Mannironi, C., Scerch, C., Fruscoloni, P., and Tocchini-Valentini, G. P. (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.