J Mol Evol (2002) 54:298–311
DOI: 10.1007/s00239-001-0045-6

JOURNAL OF MOLECULAR EVOLUTION

Phenylalanine-Binding RNAs and Genetic Code Evolution

Mali Illangasekare, Michael Yarus

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

Abstract. We isolated RNAs by selection–amplification, selecting for affinity to Phe–Sepharose and elution with free L-phenylalanine. Constant sequences did not contain Phe condons or anticodons, to avoid any possible confounding influence on initially randomized sequences. We examined the eight most frequent Phe-binding RNAs for inclusion of coding triplets. Binding sites were defined by nucleotide conservation, protection, and interference data. Together these RNAs comprise 70% of the 105 sequenced RNAs. The K_D for the strongest sites is ≈50 μM free amino acid, with strong stereoselectivity. One site strongly distinguishes free Phe from Trp and Tyr, a specificity not observed previously. In these eight Phe-binding RNAs, Phe codons are not significantly associated with Phe binding sites. However, among 21 characterized RNAs binding Phe, Tyr, Arg, and Ile, containing 1342 total nucleotides, codons are 2.7-fold more frequent within binding sites than in surrounding sequences in the same molecules. If triplets were *not* specifically related to binding sites, the probability of this distribution would be 4.8×10^{-11}. Therefore, triplet concentration within amino acid binding sites taken together is highly likely. In binding sites for Arg, Tyr, and Ile cognate codons are overrepresented. Thus Arg, Tyr, and Ile may be amino acids whose codons were assigned during an era of direct RNA–amino acid affinity. In contrast, Phe codons arguably were assigned by another criterion, perhaps during later code evolution.

Key words: Amino acid — Triplet — RNA — Selection amplification — Direct RNA template — Modification

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

Introduction

RNA binding sites for phenylalanine are of interest in light of previously selected Phe/RNA reactions. For example, there exist highly specific RNA catalysts for the synthesis of Phe–RNA (Illangasekare and Yarus 1999a) as well as RNA catalysts for the formation of peptide bonds containing Phe (Zhang and Cech 1997; Illangasekare and Yarus 1999b). Finally, RNAs activate Phe by mixed anhydride formation, paralleling modern aminoacyl RNA synthetases (Kumar and Yarus 2001). Such binding sites also bear on a question of biochemical specificity. Can RNA structures distinguish the chemically similar planar aromatic sidechains of phenylalanine, tyrosine, and tryptophan (Yarus 2000)?

Amino acid binding sites composed of RNA may also bear on the origin of the genetic code (Yarus 1998; Szathmáry 1999). The genetic code is likely to be ancient, even seen against a geological time scale embracing the full history of planet Earth. However, if amino acids and codons were once chemically linked (Woese et al. 1966), then that link should still be demonstrable today. In fact, amino acids with definitive RNA interactions (though not necessarily the persistence of the initial assignments) are required by the RNA world hypothesis. That is, coding must have appeared in an ancestral RNA world to originate the coded peptides that ultimately replaced RNA catalysts, producing succession to current nucleoprotein-based biology. Direct succession from an RNA world to current biology is itself strongly supported by the universal incorporation of varied, essential ribonucleotide cofactors into many protein catalysts (White 1976).

It is exceedingly improbable that the entire coding table quickly appeared in its modern form. Instead, the code itself evolved (reviewed by Knight et al. 1999),

with some codons perhaps assigned later by a logic independent of RNA–amino acid interaction (Freeland and Hurst 1998). Such theoretical expectations (Woese et al. 1966) are strongly reinforced by the experimental fact that the code can evolve even in modern complex genomes (Osawa et al. 1992). Such coding reassignments follow rules that have nothing to do with RNA–amino acid interaction (Yarus and Schultz 1997; Knight et al. 2001). Therefore the problem of proving a stereochemical origin, or of finding authentic initial RNA–amino acid interactions, is complicated by the probability that modern coding assignments are of mixed origin.

A possible resolution of this ambiguity was suggested by the appearance of five of six codons for arginine conserved within a natural RNA binding site for the amino acid arginine (reviewed by Yarus 1993). Coding triplets and their amino acid form a stereoselective and sidechain-specific complex within a larger RNA structure, the splicing cosubstrate (G nucleotide) site of group I self-splicing RNAs (Yarus 1988; Yarus and Christian 1989). Thus some codons, in particular those assigned in an RNA world using RNA interactions, might be nucleotide triplets extracted from the vicinity of aboriginal RNA binding sites.

The idea, in generalized form, is potentially testable. Given amino acid binding sites from selection–amplification (Famulok and Szostak 1992; Connell et al. 1993), one might be able to show that codons occur more frequently than expected within some amino acid binding sites (Knight and Landweber 1998). Binding-site nucleotides can be distinguished by nucleotide conservation in independent RNA isolates of similar sites or by chemical protection and interference (Yarus 2000) with an amino acid ligand. Codons that are associated with their RNA binding sites may then be those initially assigned by RNA–amino acid interaction, that is, in an RNA world. Here we update this argument by discussing the most frequently isolated (most probable) RNAs recovered from random ribonucleotide pools when phenylalanine binding is selected.

Methods

Preparation of the Selection Matrix

Phenylalanine was coupled to EAH Sepharose 4B (Amersham–Pharmacia) via its carbonyl group. Ten micromoles of Fmoc–phenylalanine (Bachem), 10 μmol of PyBOP [benzotriazole-1yloxytrispyrolidinophosphonium hexafluorophosphate (Novabiochem)], 20 μmol of diisopropylethylamine (DIPEA), and 30 μmol of EAH Sepharose in dimethylformamide (DMF) were combined at 24°C for 2 h as described previously (Majerfeld and Yarus 1994). After washing the Sepharose with DMF, a mixture of 900 μmol acetic anhydryde and 900 μmol DIPEA in 1-ml DMF was added to block unreacted amino groups. The resin was then washed extensively with DMF and the Fmoc protecting group was removed with 20% piperidine in DMF. The extent of coupling was determined by measuring the amount of Fmoc released, using absorption at 301 nm (ϵ_M = 9700). The concentration of Phe on EAH Sepharose was approximately 2 m*M*.

Selection Method

An initial RNA pool of $>10^{14}$ unique sequences was generated by T_7 RNA polymerase transcription of the following DNA template:

5′ taatacgactcactatagggacgacgatgacacgata $(N)_{80}$ catacgccgatcacatgacca

Selection at 24°C used a 200-μl column of Phe–Sepharose. The selection buffer was 50 m*M* Hepes, pH 7.0, 0.3 *M* NaCl, 5 m*M* $CaCl_2$, and 5 m*M* $MgCl_2$. The wash buffer contained 1 m*M* glycine, and the elution buffer had 1 m*M* Phe in addition.

RNA (1000 pmol in the first selection round, decreasing in subsequent rounds) was heated at 65°C for 5 min in water, then the salt mixture was added to get the selection buffer. The result was transferred to 24°C and kept for at least 5 min. This renatured RNA was loaded on a Phe–Sepharose column equilibrated with the selection buffer. After the addition of RNA, the column was washed with 6–20 column vol of wash buffer (increasing wash stringency as the selection progressed) before eluting RNA with elution buffer. The elution volume was 2 ml. RNA was precipitated using glycogen as carrier, reverse transcribed, and PCR amplified, and the resulting DNA was transcribed for the next selection cycle. After cycle 4, selection was preceded by counterselection on an acetylated EAH Sepharose column to remove RNA with an affinity for resin. At the 10th round of selection, eluted RNA was split into two fractions. After 2.4 ml of wash buffer the first fraction was eluted with 50 μ*M* Phe in elution buffer (1 ml), and the next fraction was eluted with 1 m*M* Phe in elution buffer as in previous cycles (1 ml). The first fraction was taken through two more cycles (see text), while only one selection cycle was performed with the second fraction before cloning. Amplified cDNA from the 12th cycle of 50 μ*M* Phe elution (pool 1) and the 11th cycle of 1 m*M* Phe elution (pool 5) was cloned (Novagen pT7 Blue-3 Perfectly Blunt cloning kit) and the individual clones were sequenced.

Chemical Modification and Enzymatic Assays

DMS (A and C modified) and CMCT (G and U modified) reactions were performed as described (Krol and Carbon 1989) except that DMS treatment was done in 50 m*M* Hepes, pH 7.0, 50 m*M* NaCl, 5 m*M* $MgCl_2$, and 5 m*M* $CaCl_2$. The CMCT reaction buffer was 50 m*M* sodium borate, ph 8.0, 50 m*M* NaCl, 5 m*M* $MgCl_2$, and 5 m*M* $CaCl_2$. In the protection experiments, about 0.5 pmol of folded RNA was incubated with or without Phe for 15 min at room temperature. Modification was initiated with the addition of DMS (0.5% final concentration; Aldrich Chemical Company) or CMCT (10 mg/ml; Fluka) and proceeded for 20 min at room temperature. The reaction was stopped by removing the modifying agent by either buffer exchange using Micro Bio-spin 30 chromatography columns (Bio-Rad) or ethanol precipitation. Modifications were identified by primer extension with AMV reverse transcriptase and denaturing PAGE and quantified using phosphorimaging (Bio-Rad GS525). The K_D for ligand binding was determined from the relationship of modification band intensity and ligand concentration as described previously (Welch et al. 1995) using the equation

$$I = I_{sat} + I_0/(1 + (A/K_D))$$

where I_{sat} is the intensity at saturating ligand and I_0 is the difference in intensity with respect to the I_{sat} in the absence of ligand. Radioactivity of test bands was normalized to that of nearby controls.

S1 nuclease protections were performed in 50 m*M* Hepes, pH 7.0,

50 m*M* NaCl, 5 m*M* $MgCl_2$, 5 m*M* $CaCl_2$, and 10 μ*M* $ZnCl_2$. After renaturation 0.25 μ*M* 5′-^{32}P-labeled RNA was incubated with or without Phe for 15 min at room temperature, then 0.1–0.2 U/μl S1 nuclease (Gibco/BRL; 2 U/μl) was added, and the mixture incubated for another 20 min. These samples were analyzed in 10% denaturing PAGE and quantified by phosphorimaging.

K_D *Measurements from Affinity Elution*

Affinity for an immobilized ligand (dissociation constant from the phenylalanine fixed on the column) was determined by elution in the absence of ligand using

$$K_c = L_c(V_n/(V_e - V_n))$$

where L_c is the concentration of affinity ligand within the bed, V_e is the median elution volume in the absence of ligand, and V_n is the median elution volume in the absence of specific affinity (randomized RNA). This dissociation constant is approximate, for example, because the accessibility of the column bound ligand is assumed.

Dissociation constants for ligands in solution were calculated as

$$K_D = L((V_{el} - V_n)/(V_e - V_{el}))$$

where L is the concentration of dissolved ligand and V_{el} is the median elution volume of the RNA in the presence of ligand (Dunn and Chaiken 1974; Connell et al. 1993; Majerfeld and Yarus 1994; Ciesiolka et al. 1996). This procedure is more likely to give a true dissociation constant for free ligand, because it is calibrated using the chromatographic behavior of the RNA without free ligand as a control.

A minimal K_D for noneluting amino acids was derived from the estimation that more than 5 ml of 1 m*M* amino acid would be required to elute the RNA (compare Fig. 3). V_e (≈5 ml here) is the median volume of elution without ligand. For noneluting ligands, $(V_e - V_{el})$ cannot be distinguished from 0 and is treated as 1 ml at maximum to estimate the minimal K_D.

Results

L-Phenylalanine with its α-amino protected was coupled to Sepharose via amide linkage to the carboxyl group, then deprotected to make the affinity resin for selection. Such coupling emulates the conserved biological form of an amino acid activated for protein synthesis, that is, with an activating group attached to its carboxyl (as in the present biological use of esterification to ribose). This also leaves the α-amino charge free as a target for RNA binding, as it would be for a free amino acid. However, we selected sidechain-specific RNA binding sites to ensure that the binding target extends beyond interaction with this positive charge and, therefore, could specify an amino acid sequence.

In addition, in these selections constant sequences that flank the 80 randomized positions (to allow for amplification) are designed to lack both the codons and the anticodons for the target amino acid. This should help ensure that the selection measures the unforced prevalence of these sequences in a binding site selected from the randomized tract.

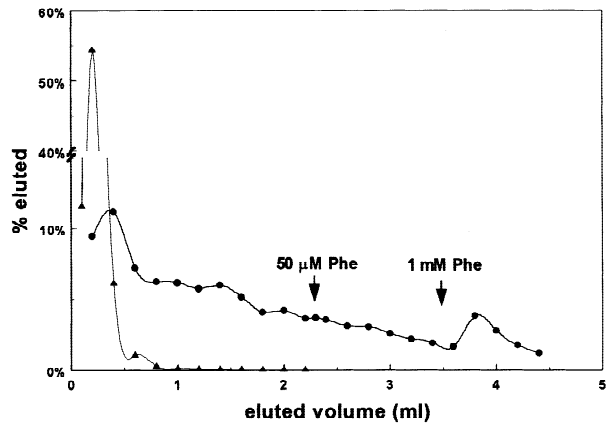

Fig. 1. Chromatographic profiles from selection: Phe elution, 10th cycle. *Circles*, Phe–Sepharose; *squares*, acetylated control column (OAc–Sepharose).

Starting with 2×10^{14} unique randomized sequences, renatured pool RNA was applied to a column of amino acid–Sepharose. Columns were washed then eluted using 1 m*M* L-phenylalanine. This protocol is intended to select RNAs that bind to the column, then are eluted with free amino acid. The resulting RNAs may therefore be expected to bind both the column ligand (analogous to an amino acid amide) and the free zwitterionic amino acid (but see below).

Phenylalanine RNAs

Figure 1 shows the elution profile from Phe–Sepharose at the 10th cycle (circles). Comparison with a control column containing acetyl groups (triangles) shows that the RNA binds specifically. That is, in the absence of phenylalanine sites, the RNA passes quickly through, and elutes quickly and quantitatively, in the column void. Selected RNAs have a clear affinity for Phe–sequence (eluting later) but give no discernible peak when eluted with 50 μ*M* Phe. Characterized sequences come primarily from the 1 m*M* Phe-eluted pool (following 3.5 ml in cycle 11; RNA clone numbers 5NN).

Additional fractionations were also performed. This included elutions with 50 μ*M* Phe, seeking RNAs with stronger binding to free amino acid. Two further cycles of selection with 50 μ*M* Phe elution were applied, then RNAs were characterized (clones numbered 1NN). Fractionation by specificity was also attempted. RNA from elution with 1 m*M* Phe at the 10th cycle (Fig. 1) was further selected with 1 m*M* Tyr (clones 2NN), 1 m*M* Trp (clones 3NN) and 1 m*M* Phe (clones 4NN). Both affinity and specificity fractionations failed to resolve a better affinity or new specificities from the 10th cycle pool, already highly selected for affinity for 1 m*M* Phe. Therefore these results are not explicitly discussed further, though some clones from these later pools were characterized (Table 1, Fig. 3).

Figure 2 shows RNAs summarized as families whose

Table 1. Sequences from the selections[a]

Family rank	Prototype sequence (Isolate No.)	Number sequenced	% of total	Specificity for some free and immobilized amino acids
1	**529**	30	29	F, not (L, N, W, Y); Fc, not (Wc, Gc, Lc, OAc)
2	**523**	10	10	F, W, Y, not (L, N); Fc, Wc, not (Gc, Lc, OAc)
3	**518**	9	9	Not F; Fc, not (Gc, Lc, OAc)
4	**530**	9	9	Not F; Fc, not (Gc, Lc, OAc)
5	**507**	7	7	Not F; Fc, not (Gc, Lc, OAc)
6	**527**	5	5	Not F; Fc, not (Gc, Lc, OAc)
7	524	5	5	Not F; Fc
8	421	4	4	Not F; Fc
9	133	4	4	—
10	**115**	2	2	Not F; Fc, not (Gc, OAc)
11	127	2	2	—
12	128	2	2	—
—	Unique	16	15	—
—	**108**	1	—	Not F; Fc, not (OAc)
—	405	1	—	Not F; Fc
—	508	1	—	Not F; Fc
—	513	1	—	Not F; Fc

[a] Families are ranked in order of descending size, or number of sequences recovered. Each family is represented by a prototype sequence, whose specificity is shown in the last column. For boldfaced sequences, binding-site nucleotides were analyzed. In the last column, F (or other single-letter amino acid abbreviations) = binds free amino acid F; F_c = binds an F-Sepharose column; not (X, Z) = the materials in parentheses or modified by "not" are not bound. In the lower section, 4 of the 16 unique sequences (from 105 RNAs cloned and sequenced) are listed.

sequences are sufficiently similar that they probably originate from the same parental molecule in the initial randomized pool. Table 1 is a census of the families, with their abundance and binding specificities.

RNAs Selected for Characterization

Figure 2 and Table 1 show that many oligoribonucleotide folds exist that satisfy our phenylalanine–Sepharose selection. From 105 sequenced molecules, we recovered binding sites originating from 28 distinct sequences in the initial randomized pool. The most prevalent binding site is that of Isolate 529 (Family 1), which alone accounts for 29% of all RNAs. This RNA was recovered in all final pools, whatever the elution protocol, even though analysis of pure RNA shows that it possesses an unprecedented specificity for free phenylalanine (below). The next most likely single sequence is the family represented by Isolate 523 (Family 2), 10% of the total clones. This has an aromatic sidechain specificity that parallels that of previously observed phenylalanine-binding RNAs (Zinnen and Yarus 1995). RNA 523 accepts all planar aromatics; that is, the binding site accommodates Phe, Tyr, and Trp. Beyond these two most frequent sites, all RNAs bound phenylalanine only on columns, though often specifically (see Families 3, 4, 5 and 6). A rationale for analysis of these column-binding RNAs appears below.

The status of one of the phenylalanine column affinities, that of Isolate 518, requires particular explanation. As shown in Fig. 2, these (Family 3; nine examples; 9% of the total) contain a decanucleotide sequence motif shared with Isolates 115 (Family 10; 2%) and 108 (a unique sequence). The threefold independent recurrence of the decanucleotide GUAGGUUCAA is unexpected for an arbitrary sequence and, therefore, presumably defines the Phe binding site. This is plausible because these molecules share their selection for Phe–resin binding and the otherwise improbable identical decanucleotide even though they are of three independent origins. This site occurred in 12 isolates, making it arguably the second most frequently selected solution to the problem of binding phenylalanine.

As can be seen from Table 1, we analyzed the phenylalanine sites of sequences that represented 73 of 105 sequenced RNAs, or 70% of all RNAs isolated. Taken together these sequences represent a substantial majority of the RNA sites that met our phenylalanine-binding selection. Therefore, we argue that they acceptably sample the RNAs that bind Phe, on the matrix and free in solution. Forty of these RNAs (39%) bound and were eluted by free phenylalanine. However, the majority had an affinity for phenylalanine only when it was bound to the matrix. These appear in substantial numbers because (being slowly eluted) they are always present as a background under a peak eluted with free amino acid (compare the front of the eluted peak in Fig. 1).

Specific Affinity for Amino Acids

Our RNAs have three types of specificity for free amino acids. A semiquantitative method to detect such affinities is to adsorb RNA to Phe–Sepharose, then elute by switching to column buffer plus test amino acid. After

Isolate number	Sequence	Number sequenced	K_C(mM)
529	GCGCGAGAAACGGUCACUAGAAUAGUGGGCCGUCAUGCUAACGCCUCUUUCGGUGUUGGGGGAAUAUUGGCCAAUCGAGU	30	0.12±0.03
523	AUUGGAUCGGUAGUAUUUAGGGUGAGACACUUCAUGCCUUUGUUGCAGGCUGGGGUGAAGGCGCUACAUGGCGUCUGAAA	10	0.12±0.04
518	GAGUGCGCCCAUUGAUAAGAGCGUAGGUUCAACGGAGUCGUAGUCGUUCUCAAGUUGACGGUGCAUUAGAGUCCU	9	0.48±0.07
530	GCCGCAAGUACGAGGGCGGCGCGCUCGACGCGGUUAGAUGAGUGUCGCCGUUUCCGAAGAUUGGCUACGCGGGG	9	1.64
507	CUGACGUUUCGAACGUCAAAUGCGGAAGUUGUCGGAGCUAAGAAACGGGUAAUGAUGCCGUUGAUGUGAAGGCGGGCUCA	7	0.86
527	CAAGCUUGCGGGUUGGCUUGUAACAAGAUCGCUACGUGUUGAACUCUGCAACACUGGUCGGAUAGCGUAAACGGG	5	0.49±0.06
524	GUCAUGUUGUGUUGUGGUCCAUCCUUUUCUACGCCAGUCGAGUCUGUAGCGUGGGAGUAAACUGAACGGUCAGACAGGA	5	0.67
421	UGUGAGGCGUUGGGGCCGCAGUGUCGUCGCCAUGGCUAACGAGGCCGGAACUAGUUUUAGCCGAAAUGAUCGAUGC	4	1.0
519	CGUAAUCUCGUACCCACUAGCAGUGAUUUGAGGCGGGGAGACAGUGUUGUAAGGUGGACGUUGUCUCGAGGGU	4	0.67
115	GAGUGCGACAAGGCAUAGCGAGAUUUGCUUUGCCGUAGGUUCAACGUUUCUUAGUUAGUCUACGCCGGAACUGGGUCUA	2	0.7±0.16
128	AACGCAGAUGUAGCUUCGUGGAGUGCCGGGGCACUGUAGGUAGGUAGGUGGUGUGCCGGAACAUGGUCGGAGA	2	
127	CGACAUGCGGGCGGUAAAUUCGUGUUGAAGUAGUUCCUCCGUUGAGUCCAGGGGGGGCUACAGAUCGGUGGCGCGGG	2	
108	CUAUGGGUAGGUUCAAGCGGGCGAAGGGAGCCUGCCAUCAGUGGCAGCUUAUUGUUCGGAAAGAGUGCCCUUGCGAGUG		0.62
405	GGUUCGCUGCAUACGGGUCCUGCAGAUGAGUGUCGUGGUCAUUCGAAUAUCACCUUAAGGGUCUGCCGAGCGAUCGGUCG		1.21
508	CCAGCACGGAACCCUGGUCGUGUCUUUCCCAAAGAUGUGACGGUUUGCACGCGGAGGUUGGGUGUUCCAUUGGUAG		1.43
513	CGUACCGCGGGGGGGGUUUUCGUGUGAGGGUCUGCUGAUGGCCCGGCGCCCAAGAGGCAGACUGGAUCGGUGA		1.0
104	GUUCACUGCGCUUGUGGGUCUGGGCAGAAUUUGAGGUCAUCAUUGAGUAAGUCGUUUAGGUGGUCGAGCGGAGACGGA		
110	GACGCGGAUGGGGUAUCGCGUUAGUGUCGAGGUGAUCGACGAUGCGCAUUUCAACACCAUGGGCCGAUGACAGG		
105	CUCGCGCUUAUCGGGAGCUGCCGGAGUGGGCACGAUCGCGGGAAGGGUCACGUUGUAUAUGGCGUUCACUGGGG		
131	ACGUAAACCGACGAGCGGUUGUGGGUGCACGCUUUGUCGGGUAGUAAAGUGUGGAAGAUGGGUCGGCAGCACGGA		
102	ACAAAAGGGCUCGUUAGUGUUUGGAAGUUUCGUCCCAUAGGAGUGACGGCUACAGAUGGAGCCGGAAUCAGGUUGAGGA		
126	GACGAGGCUUGAAUAUCAUCGUUUGGAGGAAAUAAGCCAGGUCUUUAUUUACUUUCGUGUGAGGCGGUCGGAUUGGGACU		
123	CUGAUGGAUCAAUGCGGAGGAAUGGUUCACCGAGGGUGUCUUUCGGUCAUUGUCGUUGAUCGGAGGUAUCCUGGG		
201	ACACUGCCGGGGAGGUGAGUCGUGACGGUGAUUGUGGCGUUGGGGUAGGUUAUAUGUUUCCGGGGCAAACUCGCCUUAUCU		
202	CAUGAGGCUGGAGGAGAAGUUGUAUCGGUUUGGGUUCCCGCGGAGUCAAAACGUGAGGGGGUGUAGCCUGGAGC		
208	GGCUGCACAUCGAGUCUGCAGUGUAUCAUUACAGAUCGGUUGCCCUGAACAGACAGGGUGUGCUCAGGGGAGU		
304	GAAACUCUGGGGCGAACUGCCAAGUCAUUCAGUCUUGACGCAAUUGUAGCUGCCUAGAUCGGAGAUAUGGCCGGAGAG		
423	CAUUGGCAUAGCGCGCACGGACAGCGGGAAGCGUGUAUCGCGGCGGUCAUAAAGCGCGGAUCGGUAGUAUUGUGGGUG		

Fig. 2. Prototype RNA sequences, grouped into families, with the numbers of molecules sequenced. K_C is the apparent dissociation constant from Phe–Sepharose (see Methods). *Underlined* isolate numbers are RNAs for which binding-site data were gathered.

this pulse of amino acid, the column can be checked for RNA still bound by switching to buffer plus 1 *M* NaCl with no divalents (and no amino acid). The latter condition nearly quantitatively sweeps these RNAs from the affinity resin by unfolding the binding site, which requires Mg^{2+} or Ca^{2+}.

Figures 3a–c show such experiments on the three classes of phenylalanine sites. RNA 529 (the majority RNA family) is shown in Fig. 3a. A small peak of denatured and aggregated RNA appears in the column void, at the left. The introduction of 1 m*M* L-Phe at 1.3 ml immediately produces a peak of RNA that has given up its affinity for the resin in favor of the free amino acid. The 1 *M* NaCl sweep to the right (at 2.3 ml) shows that amino acid elution of bound RNA was quantitative. Only an insignificant amount of RNA is eluted at this point (and the column showed very low Cerenkov radiation). In contrast, the introduction of 1 m*M* Tyr, Trp, Gln, or

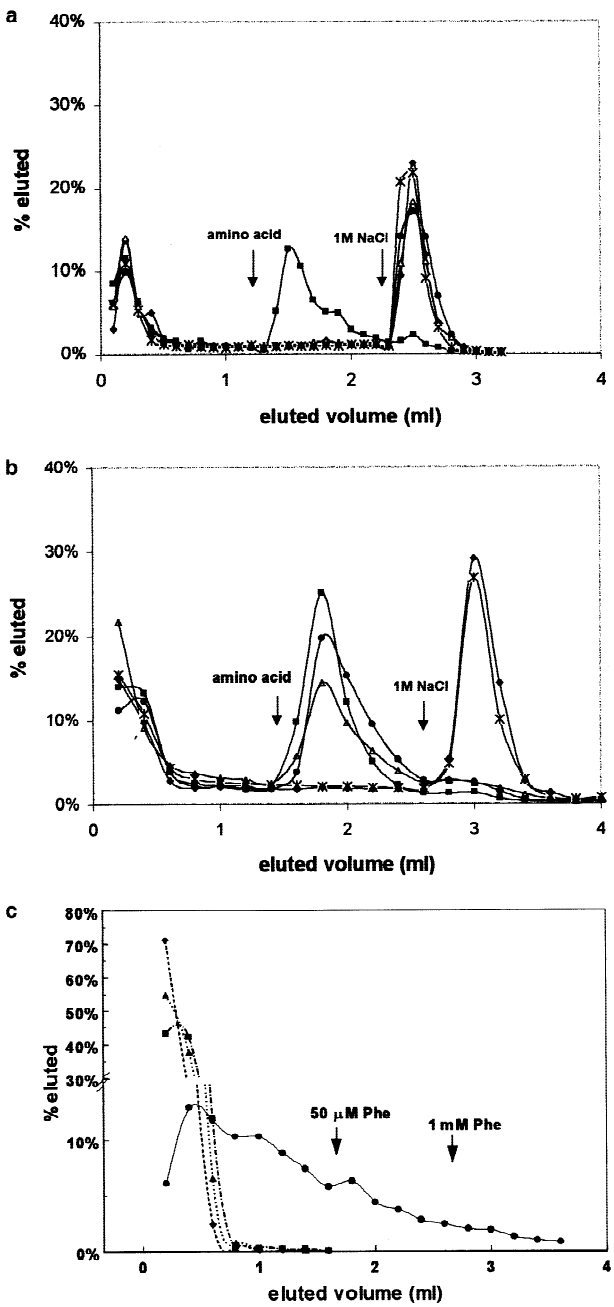

Fig. 3. Chromatography of RNA on ligand–Sepharose; elution by changing column buffer. Phe–Sepharose: *exes,* elution with 1 m*M* leucine; *filled diamonds,* elution with 1 m*M* Gln; *filled squares,* elution with 1 m*M* Phe; *filled circles,* elution with 1 m*M* Trp; *open triangles,* elution with 0.5 m*M* Tyr. **a** RNA 529. **b** RNA 523. **c** Elution of RNA 518 from ligand–Sepharose. *Filled squares,* Phe–Sepharose; *filled circles,* OAc–Sepharose; *filled diamonds,* Gly–Sepharose; *open triangles,* Leu–Sepharose.

Leu in the same way elutes no RNA, and the RNA radioactivity instead appears in the 1 *M* NaCl sweep. Therefore, Fig. 3 shows that nearly every bound molecule of RNA 529 will release Phe–Sepharose in favor of 1 m*M* free phenylalanine. Leucine and glutamine are not ligands of RNA 529 by the same test. More strikingly, the other aromatic amino acids, tyrosine and tryptophan, do not elute RNA 529 at all, even though we are testing them at about 20 times K_D for phenylalanine (see below).

The same analysis for RNA 523 (second most frequent sequence) appears in Fig. 3b. Adsorbed RNA is eluted quantitatively by 1 m*M* phenylalanine, tyrosine, and tryptophan but not by leucine or glutamine. Thus RNA 523 has the general aromatic amino acid specificity seen previously (see quantitation below). This difference has been further confirmed by chromatography on Trp–Sepharose, which is bound by RNA 523 but not RNA 529 (profile not shown).

The final kind of specificity, for RNA 518, appears in Fig. 3c, whose behavior resembles and represents that of other RNAs such as 527, 115, and 108. This RNA is not bound by control resins like acetyl–, Gly–, or Leu–Sepharose, because the RNA quantitatively appears in the void of such columns (leftward in Fig. 3c). However, the RNA binds Phe–Sepharose (Fig. 3c), showing that it has a Phe–Sepharose-specific site. However, 1 m*M* free phenylalanine does not elute, and RNA 518 also emerges from the affinity column with an unaltered profile when applied and eluted in 1 m*M* phenylalanine. Thus RNA 518 (and others with the same behavior) most likely have a phenylalanine site as a subsite within a larger amino acid–agarose site. We later consider whether such RNAs can be usefully analyzed.

Amino Acid Dissociation Constants

It is also possible to use affinity chromatography quantitatively. K_c, a dissociation constant calculated for the Phe–Sepharose column, is listed in Fig. 2. This should be considered a rough estimate, as it assumes the uniform availability of the 2 m*M* phenylalanine linked to the column. In any case, no statistically significant trend for prevalence in the pool with this quantitative measure of column affinity is evident, even if comparison is restricted to the perhaps more homogeneous Phe–Sepharose sites like RNA 518. However, it may be notable that the two RNAs that bind the free amino acid most strongly (Families 1 and 2; see below) also bind the column most strongly.

A more accurate K_D for free amino acid (rather than the column ligand) can be calculated from the displacement of the RNA toward the void volume as a constant level of free phenylalanine in the eluant increases (isocratic affinity chromatography). This K_D is normalized by the mobility in the absence of free phenylalanine, which corrects for column variables. An illustrative experiment for RNA 523, ±50 μ*M* isocratic free phenylalanine, is shown in Fig. 4a. Similar displacement toward the void by phenylalanine can be obtained for the majority of RNA 529. In contrast, RNA 518 profiles show an affinity for the resin but appear essentially unaltered when compared in buffer or 0.5 or 2 m*M* free phenylalanine (Fig. 4b). Thus these more refined measurements

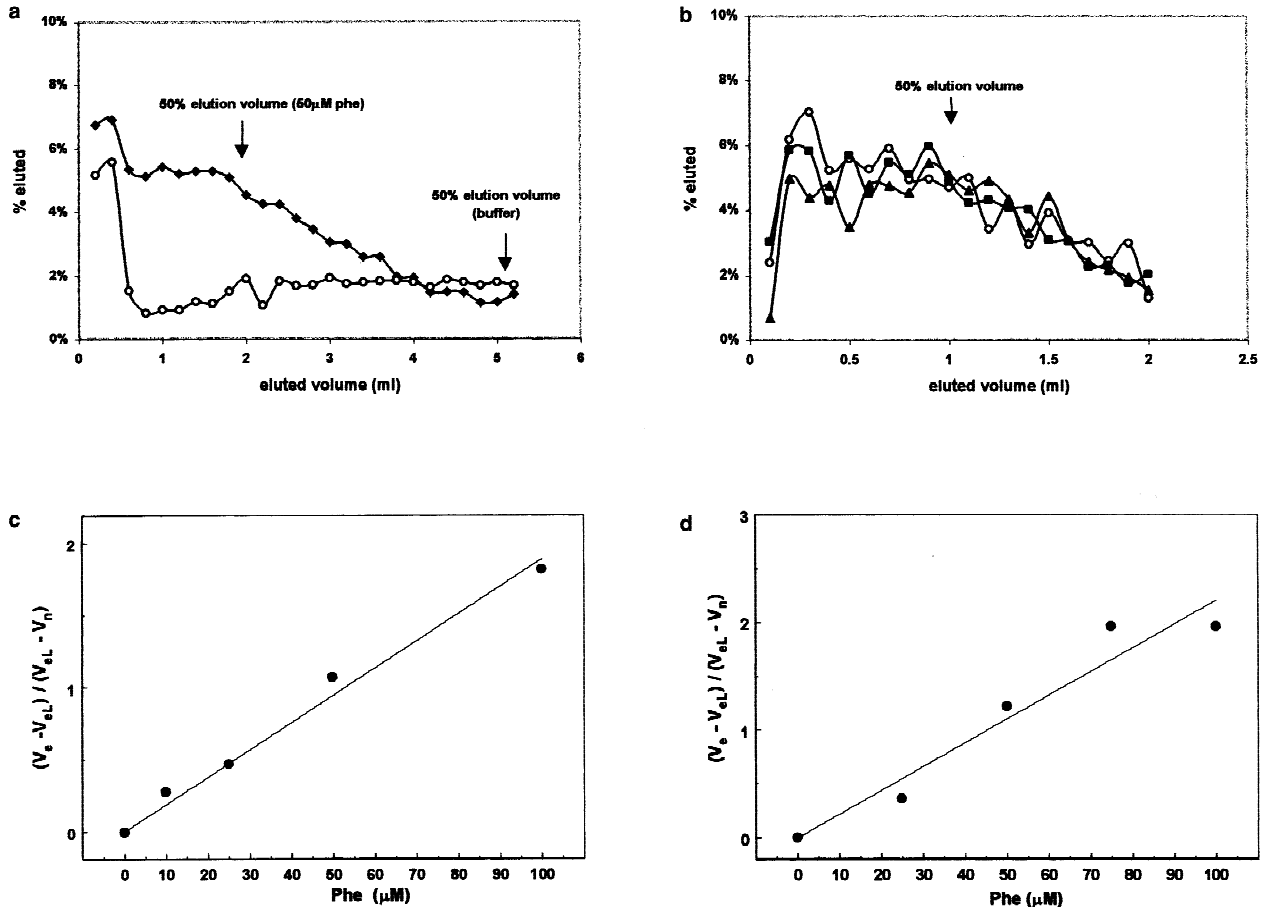

Fig. 4. Isocratic (constant eluant) affinity chromatography of RNA on Phe–Sepharose. **a** RNA 523: *open circles,* column buffer; *filled diamonds,* buffer plus 50 μ*M* Phe. **b** RNA 518: *open circles,* column buffer; *filled squares,* buffer plus 0.5 m*M* Phe; *filled triangles,* buffer plus 2 m*M* Phe. **c** Binding curve from median RNA position in isocratic chromatography, RNA 529. **d** Binding curve from median RNA position in isocratic chromatography, RNA 523.

confirm the specificities for free ligand deduced from Fig. 3 above.

Figures 4c and d show the outcome of this experiment repeated at varying concentrations of free phenylalanine. The position of the median RNA 529 molecule (Fig. 4c) not only moves to the void in free Phe but traces out the expected binding isotherm, giving a line whose slope is the reciprocal of the K_D. For RNA 529 $K_D = 55 \pm 3$ μ*M* for free phenylalanine, where the range is the standard error of the mean. In Fig. 4d, RNA 523 also reacts in the same way, suggesting that its affinity chromatography is controlled by the RNA's binding equilibrium with free phenylalanine, with $K_D = 45 \pm 7$ μ*M*.

Figure 5 shows that these K_D values can be confirmed by an independent method based on protection of the base of A nucleotides from DMS substitution by amino acid. If nucleotide modification has two intensities, higher when the binding site is unoccupied and lower when ligand is bound, then the intensity-versus-ligand concentration curve will trace an inverted binding isotherm with the lower protected intensity as a limit (Figs. 5a and b and Methods). The lower panels in Fig. 5 show gel autoradiograms, with nucleotide modification decreasing as ligand is progressively bound. Protection of A47 in RNA 523 by increasingly bound phenylalanine suggests a $K_D = 34 \pm 7$ μ*M* (Fig. 5a). The increasing protection of A55 of RNA 529 from DMS traces out a fitted binding curve (solid line) whose $K_D = 41 \pm 13$ μ*M* (Fig. 5b). Thus the K_D values from DMS protection are in reasonable agreement with, but slightly smaller than, those from quantitative isocratic affinity elution. However, given the errors of both methods, the differences are not significant. These are the most tightly binding phenylalanine sites ever isolated, with a K_D value ≈300-fold below ($\Delta\Delta G = -3.4$ kcal/mol) that of sites described previously (Zinnen and Yarus 1995). Given that these sites form several more secondary bonds to phenylalanine and can be more discriminating than earlier sites, it is perhaps not surprising that we can find no sequence similarity to the earlier oligomers.

Table 2 summarizes these binding constants, and also the K_D for the RNA 523 site binding tyrosine and tryptophan, derived by multiple isocratic affinity chromatography. These additional data show that the RNA 523 site is not casually described as nonspecific for the aromatic amino acids; it really makes virtually no distinction

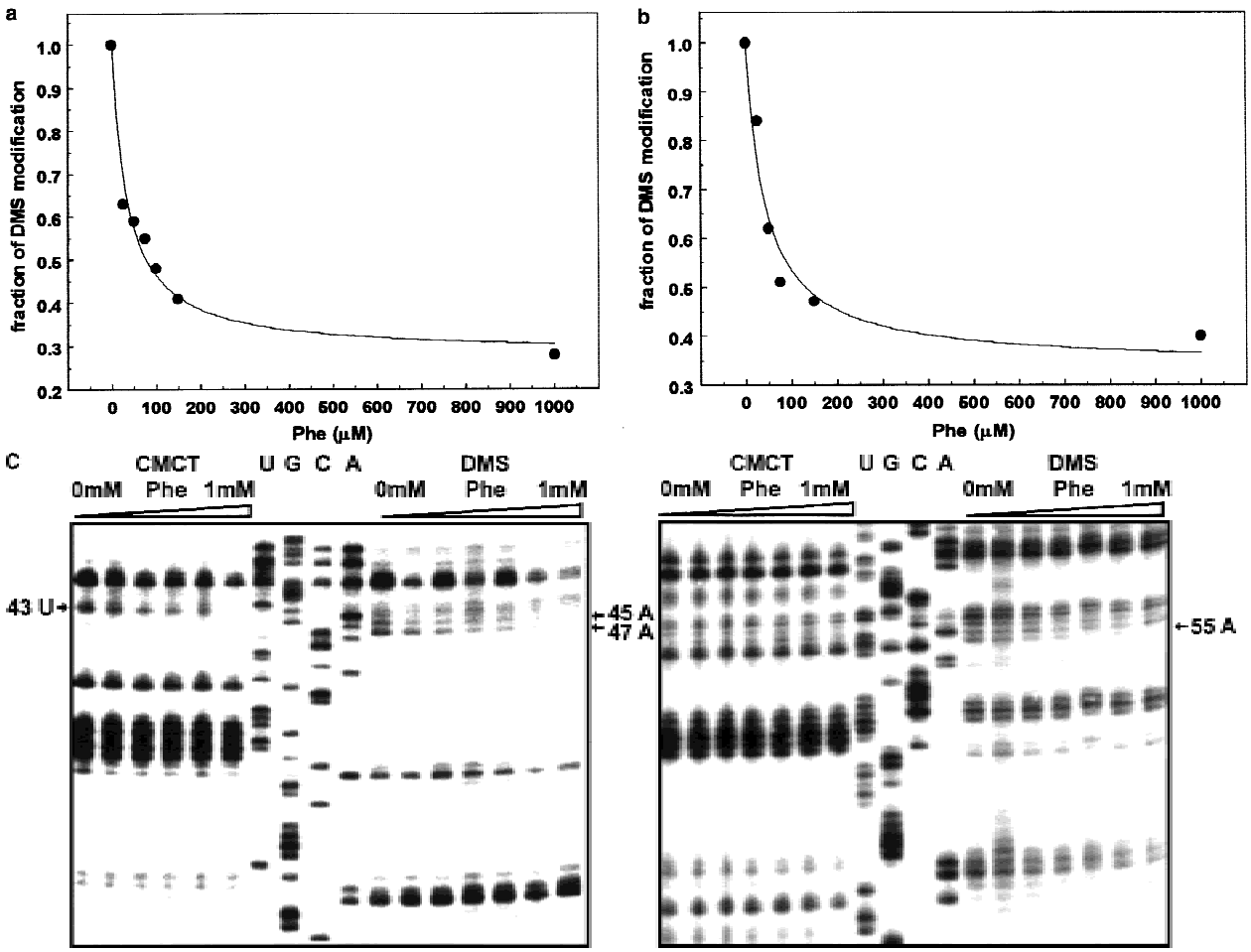

Fig. 5. K_D using DMS protections at various ligand concentrations. **a** Binding curve, A47 in RNA 523; the *line* is the least-squares binding isotherm. **b** Binding curve, A55 in RNA 529; the *line* is the least-squares binding isotherm. **c** DMS protection data in the region of A47, RNA 523. **d** DMS protection data in the region of A55, RNA 529.

Table 2. Dissociation constants for free amino acids

RNA	Ligand	K_D (μM)[a]	Method
523	L-Phe	45 ± 7	Affinity chromatography
"	L-Phe	34 ± 7	DMS protection
"	L-Phe-amide	6.7 ± 1.1*	Affinity chromatography
"	D-Phe	950 ± 50*	"
"	L-Tyr	32 ± 3*	"
"	L-Trp	51 ± 5*	"
"	L-Gln, Leu	>5000	"
529	L-Phe	55 ± 3	Affinity chromatography
"	L-Phe	41 ± 13	DMS protection
"	L-Phe-amide	2.0 ± 0.2	Affinity chromatography
"	D-Phe	14,000 ± 7000*	"
"	L-Tyr, Trp, Gln, Leu	>5000	"

[a] A superscript asterisk indicates that the range is indicated; otherwise the standard error of the mean is used when ± is indicated. See Methods for minimal K_D calculations.

among these three amino acids (compare Fig. 3). In contrast, RNA 529 must be at least moderately discriminating among aromatics. A rough minimum K_D can be calculated for amino acids that show no elution (Fig. 3). For example, free Tyr and Trp do not elute the specific RNA 529, and no RNA responds to the very different amino acid L-Leu. Minimal K_D's are implied because the percentage of RNA elution appears completely flat after 1 ml of 1 m*M* amino acid. This implies, conservatively, that the volume of eluant required to elute the median RNA is >5 ml, and the K_D > 5 m*M*. These estimates suggest that RNA 529 discriminates Phe from the other planar aromatics, Tyr and Trp, by at least 100-fold ($\Delta\Delta G \geq 2.8$ kcal/mol).

Both RNAs with an affinity for free amino acids show a greater affinity for free Phe-amide, which better resembles the column-linked amide ligand. Therefore, in both cases, the site selected likely extends to elements of the resin, even though these RNAs are quantitatively eluted by free L-Phe.

Both RNA 523 and RNA 529 are quite stereoselective, preferring the L-amino acid used during selection by 25- or 280-fold ($\Delta\Delta G = -1.9$ or -3.4 kcal/mol). Thus both RNAs whose specificity can be easily investigated (because affinity for soluble ligands can be measured) are quite selective and must make several specific molecular contacts with bound phenylalanine.

Secondary Structures Around the Amino Acid Binding Sites

To delimit the nucleotides closest to bound phenylalanine, we combined several chemical criteria. Such criteria can be shown to distinguish correctly the nucleotides proximal to an amino acid binding site where three-dimensional structure is available for comparison (Yang et al. 1996). A comparison between nucleotides implicated in the binding site and those outside the site in the same molecule is the crux of our inquiry into codon frequencies (Yarus 2000).

Figure 6 shows the calculated stable secondary folds for all eight RNAs. These secondary structures are each supported by chemical and nuclease probing (not shown). Figure 6b is also annotated with phenylalanine protection from nucleotide base modification by CMCT (G and U reactive) and DMS (A and C reactive) and phenylalanine protections from scission by nuclease S1 (all linkages potentially reactive). In addition, CMCT and DMS modification–interference with binding and elution by phenylalanine are shown. Finally, nucleotide conservations are marked with a colored background where independent isolates of the same site exist.

As an example of the data, Fig. 6a shows modification–interference data that define the three interhelix regions of the three-helix junction of RNA 523 as the Phe site. For example, the continuous tract of loop nucleotides from nt 36 to nt 47 are modified by DMS or by CMCT in RNA that flows through a Phe–Sepharose column. However, they are less modified or not modified at all in the subfraction that persists on the column and is eluted with free phenylalanine. Therefore these nucleotides must be unmodified to fold the Phe binding site. U38 is a single exception. It sometimes appears more highly CMCT modified in bound and specifically eluted RNA 523, but this band is confounded with a reverse transcriptase stop. Therefore U38 cannot be confidently classified and is conservatively placed outside the amino acid site.

Therefore we conclude that RNA 523 is likely a three-helix junction, with the phenylalanine site within the junction where interference is particularly intense. Note that all helices are supported by chemical inaccessibility, and loops, including both distal hairpins and looped junction sequences, are confirmed by accessibility. A similar helix junction model is probable for RNA 529, save that there are nucleotides involved in the binding site within the descending helical region.

Predicted helical regions frequently clearly have a more complicated fold than calculations suggest. This is likely to be true, for example, for RNA 115, where a complicated pattern of chemical susceptibilities and conservations spans a region larger than that of the calculated loops. The structures shown also ignore tertiary structure entirely, though this must be present, as in RNAs 523 and 529, where too many site nucleotides are found to all be immediate neighbors of the bound amino acid. Nucleotides are three times the size of amino acids, so space around an amino acid is quickly filled by the nucleotides of an RNA binding site. Thus we expect an unseen tertiary structure within the central loops of RNAs 523 and 529. The tertiary path within these loops defines a site (perhaps a site of intercalcation or stacking) for any planar sidechain in the former case and a highly specific site for a toluene-like sidechain in the latter RNA (perhaps a hindered site which can accept only the smallest of these amino acids).

RNA 518 has an elongated structure, in which most predicted loops and bulges are confirmed by chemical susceptibility. RNA 518 also contains a 10-nucleotide motif independently isolated in two other molecules. This is therefore probably an essential part of the phenylalanine site, and this is confirmed by the observation of both DMS and CMCT interferences in the nucleotides of the conserved sequence and at only one other nucleotide. However, this single additional nucleotide is paired to one of the conserved nucleotides. The phenylalanine site is therefore clearly within the internal loop–helix junction containing the conserved GUAGGUUCAA sequence. For the other RNAs in this family, 115 and 108, DMS and CMCT interference is also concentrated in the conserved GUAGGUUCAA sequence (Fig. 6b). These data therefore independently confirm that the decamer is the Phe binding site in all cases, as expected from its conservation in independent isolates. In calculations below, the decamer (with outside interferences, if any) is treated as the site.

RNA 527 can be approximated by the structure shown, since all secondary loop features are confirmed by chemical accessibility. However, no phenylalanine protections, S1 nuclease protections, or DMS or CMCT interferences were detectable. Thus the location of the amino acid binding site remains unknown.

To a first approximation, these phenylalanine binding sites appear to rely on local RNA structure, rather than on a more complex fold which brings together distant nucleotides.

Discussion

From a mixture of randomized ribooligomers, we isolated and characterized the major RNAs that show an affinity for the phenylalanine residue in Phe–Sepharose. These include RNAs that are quite selective for the phenylalanine sidechain or closely related molecules, and therefore might be considered for a coding system, where such distinctions are essential.

Specificity for Planar Aromatics

It has been surmised that RNAs might be capable of little distinction among phenylalanine, tyrosine, and trypto-

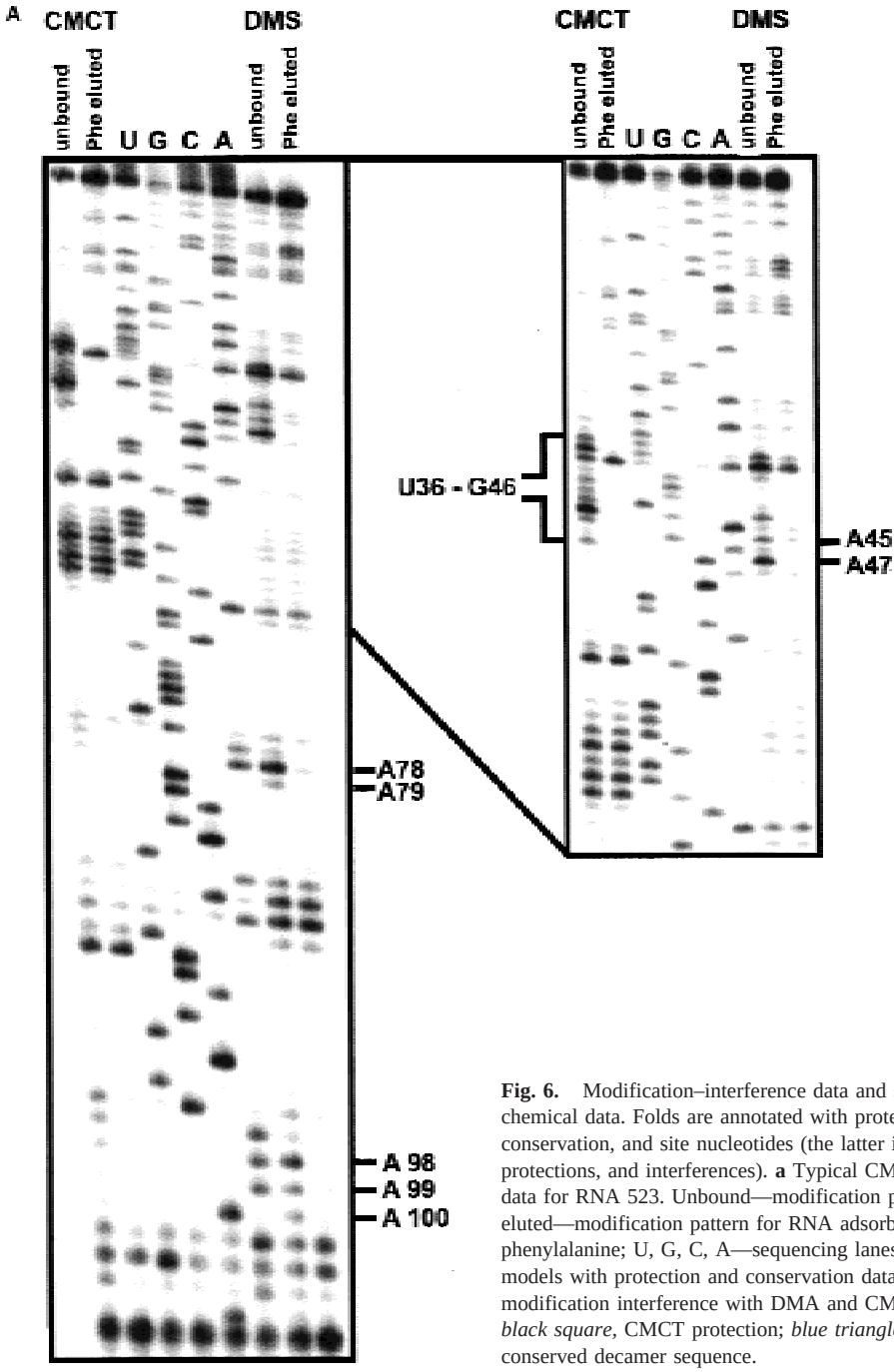

Fig. 6. Modification–interference data and summary folds, consistent with all chemical data. Folds are annotated with protections, interferences, nucleotide conservation, and site nucleotides (the latter is the sum of all conservations, protections, and interferences). **a** Typical CMCT and DMS modification–interference data for RNA 523. Unbound—modification pattern for RNA in void; Phe eluted—modification pattern for RNA adsorbed to Phe–Sepharose and eluted with free phenylalanine; U, G, C, A—sequencing lanes. **b** Approximate secondary structure models with protection and conservation data for eight RNAs. *Red circles*, modification interference with DMA and CMCT; *green ovals*, S1 nuclease protection; *black square*, CMCT protection; *blue triangles*, DMS protection; *yellow shading*, conserved decamer sequence.

phan, because these amino acids are usually all acceptable to RNA binding sites. A site selected for phenylalanine (Zinnen and Yarus 1995) or tyrosine affinity (Mannironi et al. 2000) usually has tryptophan affinity as an unselected property, perhaps because a stacking or intercalation site can similarly bind all planar sidechains. This might mean that primordial assignment of codons to aromatic amino acids via RNA affinity would be to all of them ambiguously (Yarus 2000). However, RNA 529 has phenylalanine *specificity* as an unselected property and strongly discriminates against tyrosine and tryptophan. Such a site might be hindered so that it fits only the smallest sidechain.

Therefore codons assigned based on RNA affinity need not be ambiguous among aromatics; phenylalanine at least could be specifically assigned. Tyrosine-binding RNAs have generally also bound tryptophan but moderately discriminated against phenylalanine (Mannironi et al. 2000). To complete this argument, amino acid sites in RNAs selected for tryptophan affinity, so far uncharacterized (but see Famulok and Szostak 1992), should be described.

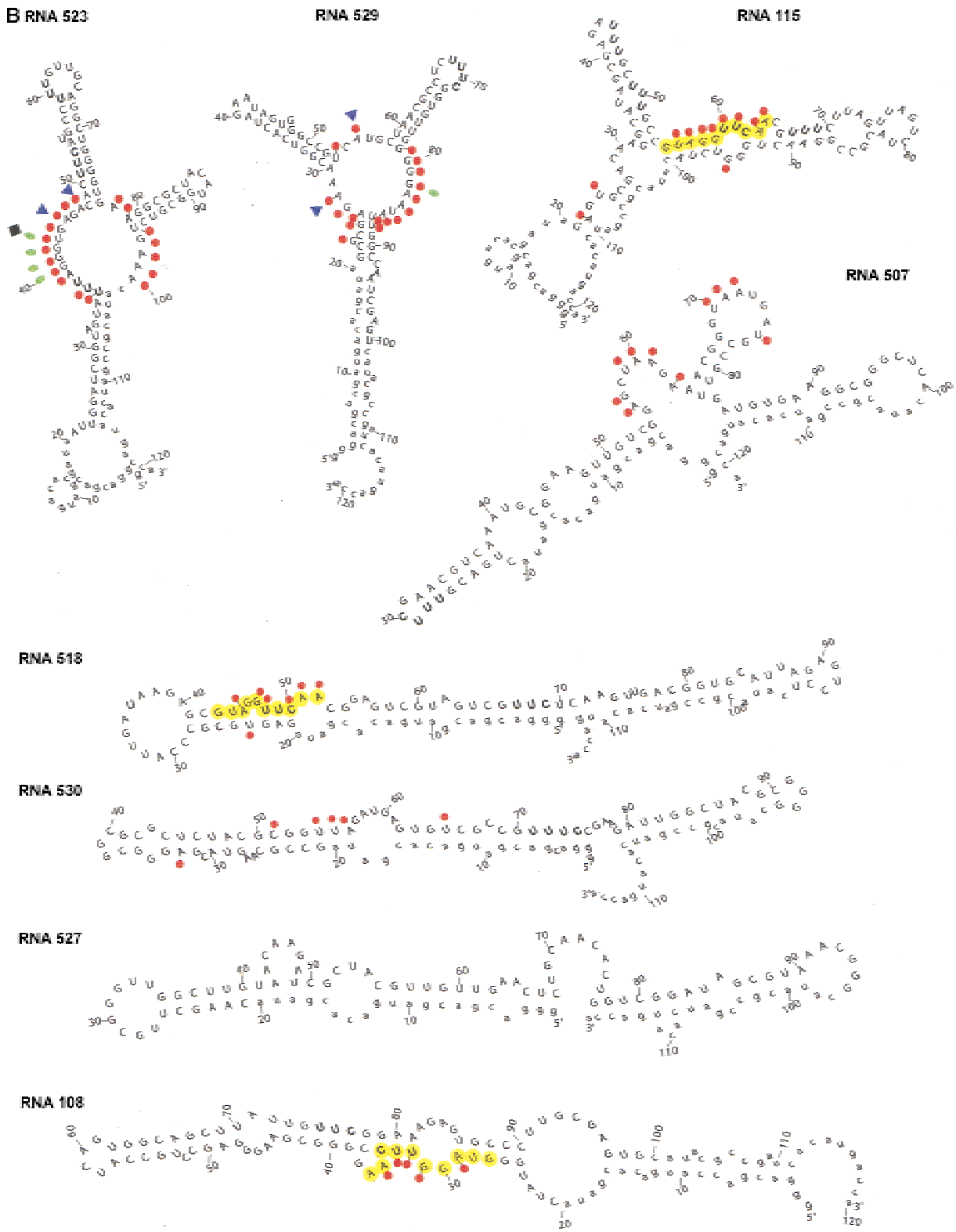

Fig. 6. Continued.

New Sites

To treat these Phe data completely, we propose to consider sites with two new properties. Therefore, what follows is in justification of these innovations, both suggested by the family of RNAs 518, 115, and 108. These RNAs share a conserved site decamer while being different elsewhere. They also bind Phe–Sepharose specifically (Table 1, Fig. 3), though (unlike RNAs 523 and 529) they are not eluted from Phe–Sepharose by free phenylalanine.

Inclusion of Nucleotides That Are Not at the Site. First, sites specific for Phe–resin (but unresponsive to elution by free phenlalanine) raise new issues. These RNA sites probably include an amino acid subsite (thereby explaining the requirement for the Phe group of Phe–Sepharose). However, they also extend to groups on Sepharose itself or, alternatively, have very stable, perhaps large site structures that are slow to release (thereby explaining why free phenylalanine does not displace RNA from resin). Thus we ask if it is valid to broaden a survey of sites, to consider binding that shows Phe specificity combined with other irrelevant structures (the postulated resin affinity and/or locking structures).

Large sites extend beyond nucleotides in direct contact with the amino acid are comparable to mistakes or overly sensitive experiments that lead us to assign too many nucleotides to the site. How does this affect the argument? There are only two possible cases: either binding sites do concentrate coding triplets or they do not. If they do, then we dilute the codons within the sites by adding irrelevant nonsite structures and thereby weaken the argument for the hypothesis by including such sites. If binding sites do not concentrate triplets, then enlarging the sites has no effect on triplet frequencies within site and nonsite sequences. Therefore, the inclusion of such compound sites, or the addition of nonsite nucleotides, will not spuriously strengthen the argument for triplets within sites. Instead this could weaken the argument. On these grounds there seems to be little danger of falsely concluding that codons are concentrated in binding sites if the sites are made too large. Therefore we can tentatively consider large sites, including those in which only some of the nucleotides approach the bound ligand.

Failure to Find All Site Nucleotides. A parallel argument applies to binding sites made too small. Experimental detection of site nucleotides will inevitably be incomplete for some RNAs. For example, in a known arginine site, a U whose uracil nucleobase is turned away from the bound ligand does not give a chemical signature when base modification is used (Yarus 2000).

Again, there are two cases; either binding sites concentrate coding triplets or they do not. If they do concentrate triplets, we tend to export triplet nucleotides to the surroundings by making the site too small. If sites do not have elevated triplet levels, then there is no effect of attributing too few nucleotides to the site. Thus we either minimize the difference between the site and its surroundings (the first case) or have no effect (the second case). Again, there seems to be little danger of falsely concluding that codons are concentrated in binding sites, this time when sites are made too small.

This is not to say that an individual contradiction cannot be devised. For example, suppose that the nucleotides mistakenly exported in a particular experiment are specifically noncodon nucleotides. Here a smaller site spuriously strengthens the argument for triplets within sites. However, the argument is that there is unlikely to be a general tendency to misassign noncodon nucleotides alone, especially across oligoribonucleotide sites of many amino acid specificities, compositions, and structures.

Thus neither type of error in enumerating site nucleotides spuriously strengthens the overall correlation between amino acid binding sites and cognate codons. Accurate site nucleotide determinations are still essential because they preserve the signal/noise ratio. Nevertheless, errors in attribution of nucleotides to sites are unlikely to yield a spurious positive conclusion, especially across a survey like the one of which the Phe binding sites are a part.

Independently and Repeatedly Isolated Sites. The second new issue concerns the occurrence of the same site in the independent molecules. The RNA 518 conserved sequence (GUAGGUUCAA; triplet underlined) recurs in three molecules of similar size, but in different positions, accompanied by broadly different sequences outside the Phe site (Fig. 2). These observations likely mean that three initial randomized RNA molecules independently gave rise to RNAs 518, 115, and 108. Any statistical evaluation of the amino acid binding site must include all independent molecules. To do otherwise amounts to choosing one set of surrounding control sequences from which to calculate triplet frequencies outside the site. Instead, it should always be more accurate to count the three independent surrounding regions together to compare with the (conserved) frequency of triplets within the site, and we do this below.

Codons in These Phe Sites. These Phe–binding RNAs are an extensive set, with eight independently isolated RNAs containing 624 total nucleotides in initially randomized positions. Taken together, the Phe-binding RNAs almost double the total number of nucleotides in fully analyzed amino acid binding RNAs from all sources (Yarus 2000). To decide if codons are concentrated in sites, we compare triplet frequency in sites and in the surrounding sequences in the same molecules. Of the 624 total nucleotides, 85 nucleotides are assigned to binding sites using conservation, protection, and interference (Fig. 6). The remaining 539 nucleotides are classed as nonsite sequences in the same molecules and supply the control estimate for triplet frequencies. There are 11/85 Phe triplet nucleotides (in UUU/UUC) among the site nucleotides and 35/539 among the nonsite nucleotides. Using the *G* test (Sokal and Rohlf 1995), the probability of this many or more triplet nucleotides within the binding sites by chance is 0.027 [p$G = 3.73$ with 1 degree of freedom (df)]. Since this probability would likely yield a positive case by chance alone if all amino acids were surveyed, it is insufficient to conclude

that UUU/UUC triplets are concentrated in phenylalanine binding sites. In addition, Phe anticodons (RAA sequences) appear in these eight binding sites with a probability of 0.088 and, therefore, are also not significantly concentrated.

We might argue instead that a better sample contains only the RNAs that bind free Phe, perhaps seeking to avoid dilution of the site with irrelevant nucleotides, as discussed above. In this case RNAs 529 (Phe specific) and 523 (aromatic specific), totaling 39% of the pool, are the test population. For these two RNAs 2 of 35 site nucleotides are triplets (529, the specific RNA, completely lacks within-site triplets), and 10 of 125 nonsite nucleotides are triplets. This codon concentration in sites has a probability of 0.37 on the hypothesis that triplet sequence and binding-site location are independent and, so, provides no support whatever for the disproportionate occurrence of UUU/UUC in phenylalanine binding sites.

Thus Phe-binding RNAs yield a negative result, with pertinent cautions about negative results. However, this outcome rests on the largest sample of amino acid sites available for any amino acid. Thus, phenylalanine is the first amino acid of the four studied which shows no significant overall tendency to concentrate its codons within selected binding site sequences. Consequently, phenylalanine is tentatively placed in the previously anticipated class of amino acids (e.g., see Di Giulio 1997) whose codons may have been assigned on a basis independent of RNA affinity, perhaps during later evolution of the code.

Status of the Hypothesis

We now briefly treat the complete data, to give Phe its proper context. In characterized amino acid binding RNAs, 28.2% (74 of 262) of nucleotides in sites are in cognate triplets, and 10.4% (112 of 1080) of nucleotides outside sites. This 2.7-fold excess of triplets within sites seems impressive, given a sample now embracing 1342 nucleotides. We can test this idea quantitatively, using Fisher's test for combined experiments (Sokal and Rohlf 1995). This allows us to combine separate results, independently obtained, to test the overarching hypothesis that triplets are not associated with amino acid sites in Arg-, Tyr-, Ile-, and Phe-binding RNAs considered together.

Combining all cases, arginine and tyrosine (Mannironi et al. 2000), the fivefold repetitively isolated specific isoleucine site [(Majerfeld and Yarus 1998); note that only one independently isolated RNA was previously used (Yarus 2000)] and the complete set of Phe RNAs (Fig. 6) a χ^2 using the Fisher test yields of 65.02 (8 df). This corresponds to a probability of 4.8×10^{-11} that coding triplets and binding sites are unrelated in the complete data. In other words, we would expect about one similar or more positive localization in 20 billion trials with a truly random triplet distribution. Thus, despite the (now realized) expectation that some amino acids will not show the predicted tendency, available data are cumulatively strongly in favor of the concentration of coding triplets within newly selected amino acid binding sites.

Table 3. Fisher's test for combined experiments

Amino acid [No. RNAs]	Probability (P)	$-2\ln P$[a]
Arginine [5]	8.12×10^{-8}	32.7
Isoleucine [5]	6.46×10^{-4}	14.7
Tyrosine [3]	5.43×10^{-3}	10.4
Phenylalanine [8]	2.67×10^{-2}	7.25

[a] $-2\sum\ln P = 65.0$, which is χ^2 distributed, 8 df $\Rightarrow P = 4.8 \times 10^{-11}$.

Further, sufficient data now exist to partition experiments in many ways, still leaving a significant argument. For example, does evidence for triplet concentration exist outside the amino acid arginine? We can omit the highly significant arginine sites. For isoleucine, tyrosine, and phenylalanine only, from Table 3 the probability of the observed codon distribution is 1.4×10^{-5} ($\chi^2 = 32.4$ with 6 df). Accordingly, we have substantial reason to believe that codons are concentrated in binding sites, even without consideration of arginine.

Criticism of Triplet Localization

This kind of argument for triplet concentration has been criticized, based on the initial arginine aptamer data (Ellington et al. 2000). There are three principal arguments. First, one can choose sets of oligomers for which codon concentration in arginine binding sites is not significant. However, it is incorrect to choose specific sequences to disprove these ideas, just as it would be incorrect to select sequences that support them. The prediction to be tested is rather that codons, while perhaps present in some sites and absent in others, will be overrepresented overall among amino acid binding sites, without preselection of compositions or sequences. Here, as elsewhere, we have usually used the most likely binding sites for our test. However, the overrepresentation of arginine triplets in arginine binding sites is exceedingly robust to varied means of choosing test sequences, even if sites are guessed on the basis of homology rather than experimentally determined (Knight and Landweber 2000). Second, it is argued that concentrations of individual codon nucleotides in binding sites are not statistically significant. While this may be true, it is not relevant. The prediction is that triplet *sequences* will be elevated within binding sites. Statistical tests that discard all nucleotide sequence information therefore cannot test the hypothesis. Finally, arginine codons are not concentrated in ribooligomers that bind arginine-rich peptides [rather than free arginine (Ellington et al. 2000)]. These experiments suffer from the fact that sites and surrounds are not

distinguished, reducing the ability to see the predicted effect (as in the above discussion of sites made too large). Further, control triplet frequencies are calculated from composition and a second hypothesis, rather than measured experimentally. In addition, free arginine is a different chemical object, with new constraints such as added attraction (α-amino) and repulsion (α-carboxyl) for RNA. Finally, peptides introduce the uncertain effects of neighboring amino acids which must also be accommodated if arginine is bound. However, perhaps the absence of codons in such Rev and Rex peptide binding sites could be taken as evidence against codon assignment via a primordial peptide ligase (Ellington et al. 2000). For these reasons, and considering the extension of the argument beyond arginine, these objections do not seem compelling.

Acknowledgments. We thank Ico de Zwart, Alexendre Vlassov, Vasant Jadhav, and Irene Majerfeld for useful suggestions on the draft manuscript, Rob Knight for anticodon calculations, and the USPHS (Research Grants GM 30881 and GM 48080) for support.

References

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

Di Giulio M (1997) On the origin of the genetic code. J Theor Biol 187:573–581

Dunn M, Chaiken IM (1974) Quantitative affinity chromatography. Determination of binding constants by elution with competitive inhibitors. Proc Natl Acad Sci USA 71:2382–2385

Ellington AD, Khrapov M, Shaw CA (2000) The scene of a frozen accident. RNA 6:485–498

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

Freeland SJ, Hurst DL (1998) The genetic code is one in a million. J Mol Evol 47:238–248

Illangasekare M, Yarus M (1999a) Specific, rapid synthesis of phe-RNA by RNA. Proc Nat Acad Sci USA 96:5470–5475

Illangasekare M, Yarus M (1999b) A tiny RNA that catalyzes both aminoacyl-RNA and peptidyl-RNA synthesis. RNA 5:1482–1489

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

Knight RD, Freeland SJ, Landweber LF (1999) Selection, history and chemistry: The three faces of the genetic code. TIBS 24:241–247

Knight RD, Landweber LF, Yarus M (2001) How mitochondria redefine the code. J Mol Evol 53:299–313

Krol A, Carbon PA (1989) A guide for probing native and small nuclear RNA and ribonucleoprotein structures. Methods Enzymol 180:212–227

Kumar RK, Yarus M (2001) RNA-catalyzed amino acid activation. Biochemistry 40:6998–7004

Majerfeld I, Yarus M (1994) An RNA pocket for an aliphatic hydrophobe. Nature Struct Biol 1:287–292

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

Mannironi C, Scherch C, Fruscoloni P, Tocchini-Valentini GP (2000) Molecular recognition of amino acids by RNA aptamers: The evolution into a L-tyrosine binder of a dopamine binding motif. RNA 6:520–527

Osawa S, Jukes TH, Watanabe K, Muto A (1992) Recent evidence for the evolution of the genetic code. Microbiol Rev 56:229–264

Sokal RR, Rohlf FJ (1995) Biometry: The principles and practice of statistics in biological research, 3rd ed. WH Freeman, New York

Szathmáry E (1999) The origin of the genetic code: Amino acids as cofactors in an RNA world. Trends Genet 15:223–229

Welch M, Chastang J, Yarus M (1995) An inhibitor of ribosomal peptidyl transferase using transition-state analogy. Biochemsitry 34:385–390

White III HB (1976) Coenzymes as fossils of an earlier metabolic state. J Mol Evol 7:101–104

Woese CR, Dugre DH, Dugre AS, Kondo M, Saxinger WC (1966) On the fundamental nature and evolution of the genetic code. Cold Spring Harbor Symp Quant Biol 31:723–736

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

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

Yarus M (1993) An RNA-amino acid affinity. In: Gesteland RF, Atkins JR (eds) The RNA world. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

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 (2000) RNA-ligand chemistry: A testable source for the genetic code. RNA 6:475–484

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

Yarus M, Schultz DW (1997) Response: Further comments on codon reassignment. J Mol Evol 45:1–8

Zhang B, Cech TR (1997) Peptide bond formation by *in vitro* selected ribozymes. Nature 390:96–100

Zinnen S, Yarus M (1995) An RNA pocket for the planar aromatic side chains of phenylalanine and tyrosine. Nucleic Acids Symp Ser 33:148–151