A Model of Proto-Anti-Codon RNA Enzymes Requiring L-Amino Acid Homochirality

Albert Erives

Abstract All living organisms encode the 20 natural amino acid units of polypeptides using a universal scheme of triplet nucleotide "codons". Disparate features of this codon scheme are potentially informative of early molecular evolution: (i) the absence of any codons for D-amino acids; (ii) the odd combination of alternate codon patterns for some amino acids; (iii) the confinement of synonymous positions to a codon's third nucleotide; (iv) the use of 20 specific amino acids rather than a number closer to the full coding potential of 64; and (v) the evolutionary relationship of patterns in stop codons to amino acid codons. Here I propose a model for an ancestral *proto-anti-codon RNA* (pacRNA) auto-aminoacylation system and show that pacRNAs would naturally manifest features of the codon table. I show that pacRNAs could implement all the steps for auto-aminoacylation: amino acid coordination, intermediate activation of the amino acid by the 5′-end of the pacRNA, and 3′-aminoacylation of the pacRNA. The anti-codon cradles of pacRNAs would have been able to recognize and coordinate only a small number of L-amino acids via hydrogen bonding. A need for proper spatial coordination would have limited the number of chargeable amino acids for all anti-codon sequences, in addition to making some anti-codon sequences unsuitable. Thus, the pacRNA model implies that the idiosyncrasies of the anti-codon table and L-amino acid homochirality co-evolved during a single evolutionary period. These results further imply that early life consisted of an aminoacylated RNA world with a richer enzymatic potential than ribonucleotides alone.

A. Erives (✉)
Department of Biological Sciences, Dartmouth College,
Hanover, NH 03755, USA
e-mail: Albert.J.Erives@Dartmouth.edu

Introduction

The correspondence between codons and specific amino acids, i.e., the codon table, is a universal feature of the ribosomal translation of mRNAs into polypeptide sequences. By advancing our understanding of the evolutionary origin of amino acid codons, we may be able to advance our understanding of the ancestors of all living organisms (Woese et al. 1966; Maizels and Weiner 1994; Weiner and Maizels 1999; Marck and Grosjean 2002). For this task, the following disparate, unexplained features of this codon scheme are potentially informative:

(i)   the absence of any codons for D-amino acids,
(ii) alternate codon patterns for some amino acids (e.g., 5′-CGN and 5′-AGR for L-Arg),
(iii) confinement of synonymous positions to a codon's third nucleotide,
(iv) specification of 20 amino acids despite a coding potential of 64 amino acids, and
(v)   relation of stop codons to amino acid codons.

As the tRNA molecules bear the anti-codon sequences that recognize codons, tRNAs must be considered in the task to decode codon evolution. Whole-genome sequence assemblies for archaebacteria have revealed new aspects of the ancestral tRNA genes. In addition to canonical tRNAs, which are encoded by a single exon and form a complete cloverleaf secondary structure, archaeal genomic tRNA repertoires include tRNA genes with alternate structures. Some tRNA genes contain a single intron immediately 3′ to the anti-codon, thus defining separate 5′ and 3′ exonic

portions of the tRNA molecule (Sugahara et al. 2006; Sugahara et al. 2007). In addition, some tRNA molecules are produced by the ligation of RNAs encoded by separate genes for these same 5′ or 3′ tRNA moieties (Randau et al. 2005). These "unjoined" tRNA "half-mers" occur throughout the phylogenetic tree for tRNAs and are assumed to represent the ancestral structure. This conclusion is also supported by the observation that the cloverleaf tRNA structure of modern tRNAs can be recreated by a head-to-tail dimer of a "proto-tRNA" half-mer (Di Giulio 2006).

Dimers of proto-tRNA half-mers would be characterized by a novel pairing of anti-codon and amino acid acceptor tail at each of the two ends of the dimer (Di Giulio 2006; Fujishima et al. 2008). This novel proximity of an anti-codon to an amino acid acceptor tail would also occur in the single hairpin form. Subsequent tandem duplications of some half-mer tRNA genes, followed by gain of splice sites or loss of the intergenic gap would then result in the modern repertoire of tRNA genic structures. This repertoire would include the extant tRNA molecules in which the anti-codon and acceptor stem are located at extreme opposite ends of the tRNA molecule (Di Giulio 2006).

Here I show that the physical proximity of anti-codon and acceptor stem in ancestral tRNAs is relevant to a long-sought goal of deriving amino acid/codon pairing rules from an ancestral nucleotide-based receptor-ligand recognition system (Woese et al. 1966). I propose a structural model of anti-codons as a stereochemical ligand coordinating pocket in short, hair-pinned, *proto-anti-codon RNAs* (pacRNAs). PacRNAs resemble Hopfield's hairpin tRNA precursors, which he postulated over 30 years ago without any details as to their chemistry (Hopfield 1978). I show that the pacRNA anti-codon sequence $5'\text{-}(N_1)N_2N_3U_4$ is limited to coordinating only certain amino acids in the orientation required for aminoacylation of the adenosine that is base-paired to U_4. As such, the pacRNA molecule constitutes a viable receptor coordinating system with high specificity for specific amino acid ligands. I also show how the pacRNA molecule may have accomplished both the intermediate step of activating the amino acid carboxyl group and the final 3′-aminoacylation. Given D-ribose nucleotide chirality, this aminoacylation coordination system would operate only on the levorotary chiral isoforms of branched amino acids (L-amino acids) but not their dextrorotary isoforms (D-amino acids). However, any initial surplus of L-amino acids will have initiated a preference for D-ribose nucleic acid chemistry, which would then have reinforced the usage of L-amino acids in an evolutionary bootstrap process. Thus, regardless of causal directionality, the pacRNA model marries the questions on the evolutionary origins of ribose and amino acid homochiralities as complementary aspects of one phenomenon.

Results

The pacRNA Model Links Ribose Chirality to Amino Acid Chirality

I describe an RNA-based amino acid coordination and auto-aminoacylation system in which proto-anti-codon "cradle" sequences are sandwiched between 5′-stem-loop ceilings and 3′-aminoacylation acceptor stems (Fig. 1a and Materials and Methods). This architecture could have evolved within any nucleic acid molecular system capable of forming a hairpin followed by an exposed anti-codon sequence and an adjacent stem-looped acceptor stem. Because many such molecular lineages may have evolved that were unrelated to the proto-tRNA lineage, I refer to all of them generically as *proto-anti-codon RNAs* (pacRNAs, see Fig. 1a). For simplicity of exposition, I assume the presence of a D-ribose chirality. Nonetheless, it is important to note that the model allows chiral preferences in either the ribose sugar or amino acids to bias chiral preferences in the other. Thus, an initial surplus of L-amino acids could have initiated a preference for D-ribose nucleic acids,
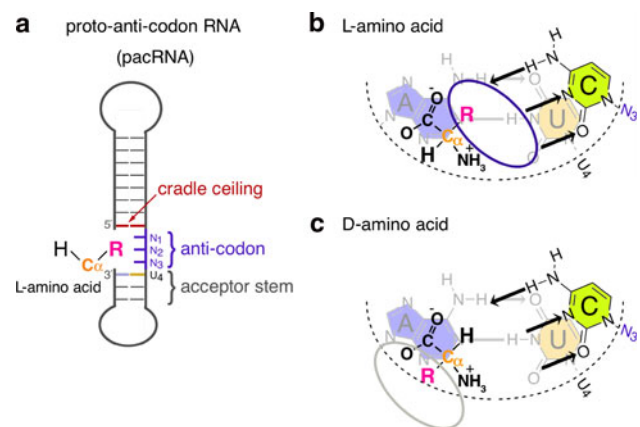

Fig. 1 proto-anti-codon RNAs. **a** Relevant features of the *proto-anti-codon RNA* (pacRNA): steric occluding ceiling made by the first base-pair of the hairpin roof (*red* paired bases) over an anti-codon pocket (*purple* unpaired bases), an acceptor stem target sequence, and the key adenosine nucleotide to be aminoacylated (*faint blue*) by an L-amino acid, which has its R-group side chain facing into the anti-codon surface. **b** and **c** Shown are *top–down* view of the bottom two nucleotides shown in **a** and an amino acid zwitterion, which is placed in a certain fixed orientation required for 3′-aminoacylation of the adenosine. **b** An L-amino acid with its R-group side chain facing into the pacRNA molecule where it can interact with the anti-codon nucleoside bases. **c** A D-amino acid with its R-group side chain facing away from the pacRNA molecule. The adenine:uracil base-pair (*faint blue* and *yellow* nucleotides) is located underneath the cytosine nucleotide (*green base*), which is linked immediately 5′ above the uracil. The Cα atom (*orange*) of the amino acid rises to the level of the depicted cytosine base, which in the proposed model corresponds to nucleotide N_3 of an anti-codon sequence. Atoms connecting ribose C1′ to the carbonyl oxygen are not shown

which would have reinforced the preference for L-amino acids in an evolutionary bootstrap process, or vice versa (see Discussion).

A pacRNA resembles Hopfield's hairpin tRNA precursor (Hopfield 1978) except that the adjacent acceptor strand also forms a hairpin, which fixes in space the position of the 3′-end (Fig. 1a). The aminoacylated adenosine nucleotide at the 3′-end may have been part of the pacRNA acceptor stem sequence, or ligated to the pacRNA like the 5′-CCA acceptor sequence is added to extant tRNAs (Deutscher 1982; Pan et al. 2010).

I consider complementary ligands for the exposed binding pockets of pacRNAs under a specific constraint of "proper coordination". Proper coordination occurs when the physiologically dipolar form of an amino acid (a zwitterion) occupies the position and orientation necessary for activation and charging to the adenine base paired to U_4 (Fig. 1a, b). In its charging position, the $C\alpha$ atom of an amino acid rises to just about level with the nucleotide base N_3 of the anti-codon (Fig. 1a, b). If the amino acid is an L-amino acid, its branched side chain will face into the pacRNA and interact with the nucleotide bases of the anti-codon sequence (Fig. 1b). If the amino acid is a D-amino acid, its branched side chain will extend away from the pacRNA molecule entirely (Fig. 1c). I therefore propose that pacRNA stereochemistry constitutes the evolutionary origin for universal L-amino acid chirality (with the stated caveat of reciprocal influences between ribose and amino acids). Preference for L-amino acids by D-ribose nucleic

acids would not have been affected by helical handedness because this chiral preference arises from base complementarity, which is preserved under both left- and right-handed helices.

Importance of pacRNA Coordination for Catalysis

Once a pacRNA coordinated its specific amino acid (Fig. 2a), canonical activation of the amino acid carboxyl group and subsequent 3′-charging would occur within the context of the pacRNA molecule itself. A version of the suggested chemistry has already been identified in selected RNA molecules and studied (Kumar and Yarus 2001). The intermediate activation of the amino acid by the 5′-end of the pacRNA (Fig. 2b) and the subsequent 3′ aminoacylation of the pacRNA (Fig. 2c) are identical for all amino acid ligands. Nonetheless, these last two steps could have been facilitated by additional enzymatic co-factors and/or mechanisms, which the present pacRNA model does not attempt to address. Thus, the extent to which the pacRNA model allows 1-to-1 correspondence between anti-codon sequences and unique amino acid ligands (Fig. 1b) will determine the extent to which the model is useful. For this reason, the strength of the model rests predominantly with successful verification of the extent of 1-to-1 correspondence in the coordination position. Here, the proposed 3′-aminoacylation step features in this task only in informing the proper coordination to test for correspondence. (Note:
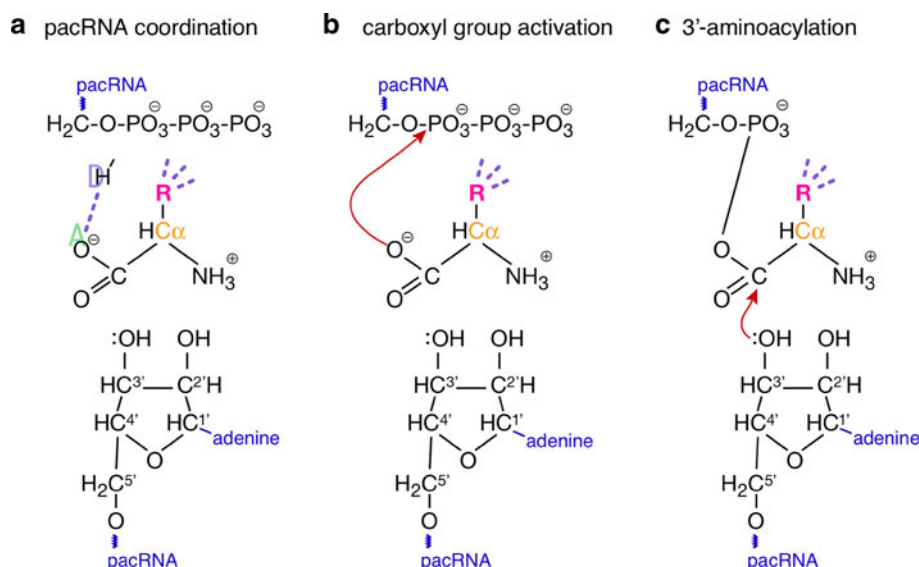

Fig. 2 pacRNA cradle chemistry. **a–c** Shown are the three steps involved in pacRNA cradle chemistry: **a** amino acid binding and coordination, **b** intermediate activation of the amino acid carboxyl group by the 5′-end of the pacRNA, and **c** 3′-aminoacylation. These diagrams depict the phosphate-ribose backbone of the adenosine nucleotide that will form a 3′-L-aminoacyladenylate molecule. This adenosine is base paired with U_4 and initiates the nucleophilic attack (*red arrow*) of the carboxyl carbon of the amino acid. This carboxyl carbon is coordinated by a H-bond (*thick purple dashes*) with the anti-codon surface. The $C\alpha$ atom (*orange*) of the amino acid rises to the level of the depicted cytosine base, which in the proposed model corresponds to nucleotide N_3 of an anti-codon sequence

Charging of modern day tRNAs by protein aminoacyl-transferases occurs via the 3'-OH or 2'-OH ribose groups but then equilibrates to the 3'-OH group non-enzymatically. Here, I do not consider 2'-aminoacylation for two reasons. First, the 2'-OH group is tilted away from the anti-codon sequence while the 3'-OH group is tilted toward it in a more accessible position. Second, the 3'-OH group is located closer to the axial H-bond donor groups of the base than the 2'-OH group.)

As in nucleotide-base complementarity, the only chemical bonding phenomenon capable of ligand coordination is hydrogen bonding. (For example, the amino acid molecules are too small to be stabilized significantly by Van Der Waals interaction.) Enzymatic catalysis of the aminoacylation reaction by pacRNAs requires only some coordinated hydrogen bonding for it to begin to function as a catalyst. Unlike double-stranded nucleotide base-pair complementarity, pacRNA hydrogen bonding does not have to hold its ligand stably together for very long for it to function as a catalyst. Thus, every additional hydrogen bond would be expected to increase catalytic potential for auto-aminoacylation.

I infer two basic principles were operative to have allowed a coherent anti-codon amino-acid receptor/charging system to emerge. The first principle is that pacRNAs used the stem duplex as steric hindrance 5' of the anti-codon sequence in order to preclude binding of longer chained competitors and thereby increase specificity. The second principle is that correct orientation of the amino acid is required for charging, thus reducing the number of amino acids with "chargeable" binding configurations. For example, in almost all cases an anti-codon sequence possesses a more extensive ensemble of binding configurations for its cognate amino acid than for any potential amino acid competitor. Furthermore, the ensemble binding states of the cognate ligand are related to each other by simple translations or rotations. These coherent ensemble states thus determine specificity for binding and positioning a unique amino acid by limiting the number of reversible binding antagonists (non-charging competitors) and virtually eliminating irreversible charging antagonists.

Because specificity of coordination is central to a working pacRNA model, I concentrate predominantly on the complementary H-bonding acceptor/donor (A/D) profiles between anti-codons and their amino acid cognates. I summarize the modeled ligand-coordinating capacities of pacRNAs with a graphic representation of their hydrogen bonding potentials (Fig. 2a). This representation depicts the Watson–Crick hydrogen-bond D and A atoms across the axial, medial, and distal columns of the cylindrical radius, as illustrated for the four possible nucleotide base pairs (Fig. 3a). These presentations summarize binding ensembles that were modeled computationally and physically and can be visualized by swinging the two tables together to find matching A and D atoms (curved arrow in Fig. 3a, b; see Materials and Methods).

pacRNAs for Short-Chained R-Groups

In the pacRNA model, L-amino acids with increasingly longer side chains interact with increasingly distant nucleotide bases located 5' of N_3 while remaining in the fixed chargeable position at N_3. This would explain why all amino acids with the shortest side-chains use di-nucleotides rather than triplet anti-codons (Figs. 3, 4). For example, the six amino acids with the shortest side groups use doublet anti-codons. These correspond to Gly (5'-XCC), Ala (5'-XGC), Pro (5'-XGG), Val (5'-XAC), Thr (5'-XGT), and Ser (5'-XGA), where "X" designates an unavailable base-paired nucleotide, i.e., a base pair ceiling. This nucleotide position is the most distant from $C\alpha$ and corresponds to the nucleotide that base-pairs with a triplet codon's third nucleotide, which is the site of synonymous positions. Thus, I propose that intrinsic pacRNA stereochemistry explains how synonymous positions in codons arose exclusively at the third position rather than being randomly distributed as expected in an arbitrary coding scheme.

Anti-Codon 5'-$X_1S_2C_3$, Gly and L-Ala

I propose that double cytosines in a 5'-$X_1C_2C_3$ anti-codon sequence can form four hydrogen bonds with glycine (Gly) (Fig. 3e–g) at multiple positions that lead to the inferred chargeable position by simple rotation (Fig. 3e). Other binding states are possible that allow Gly but not any other amino acid (Fig. 3h). I infer that the stem helix would begin at the N_1 position of the anti-codon to preclude binding by other amino acids at the receded double cytosine wall of the anti-codon pocket. L-alanine (L-Ala) might bind the third cytosine with two hydrogen bonds but its additional methyl group would prevent the more stable binding states characterized by four hydrogen bonds (Fig. 3i). I propose that the four-hydrogen bound states for Gly facilitate movement into the chargeable position, and that the single chargeable binding state for L-Ala makes it an insignificant competitor in comparison. I find similar coherent entry binding pathways for other anti-codon/amino acid relationships as well, and tentatively suggest that this constitutes a third principle feature of pacRNAs.

By replacing the second cytosine of the Gly anti-codon 5'-$X_1C_2C_3$ with guanine, the purine ring center can form an umbrella over the single methyl group of L-Ala (Fig. 4a, b). I propose that this gives it higher affinity for L-Ala versus Gly via hydrophobic packing, while precluding other amino acids given the base pair hindrance at N_1 and the purine ceiling at N_2.
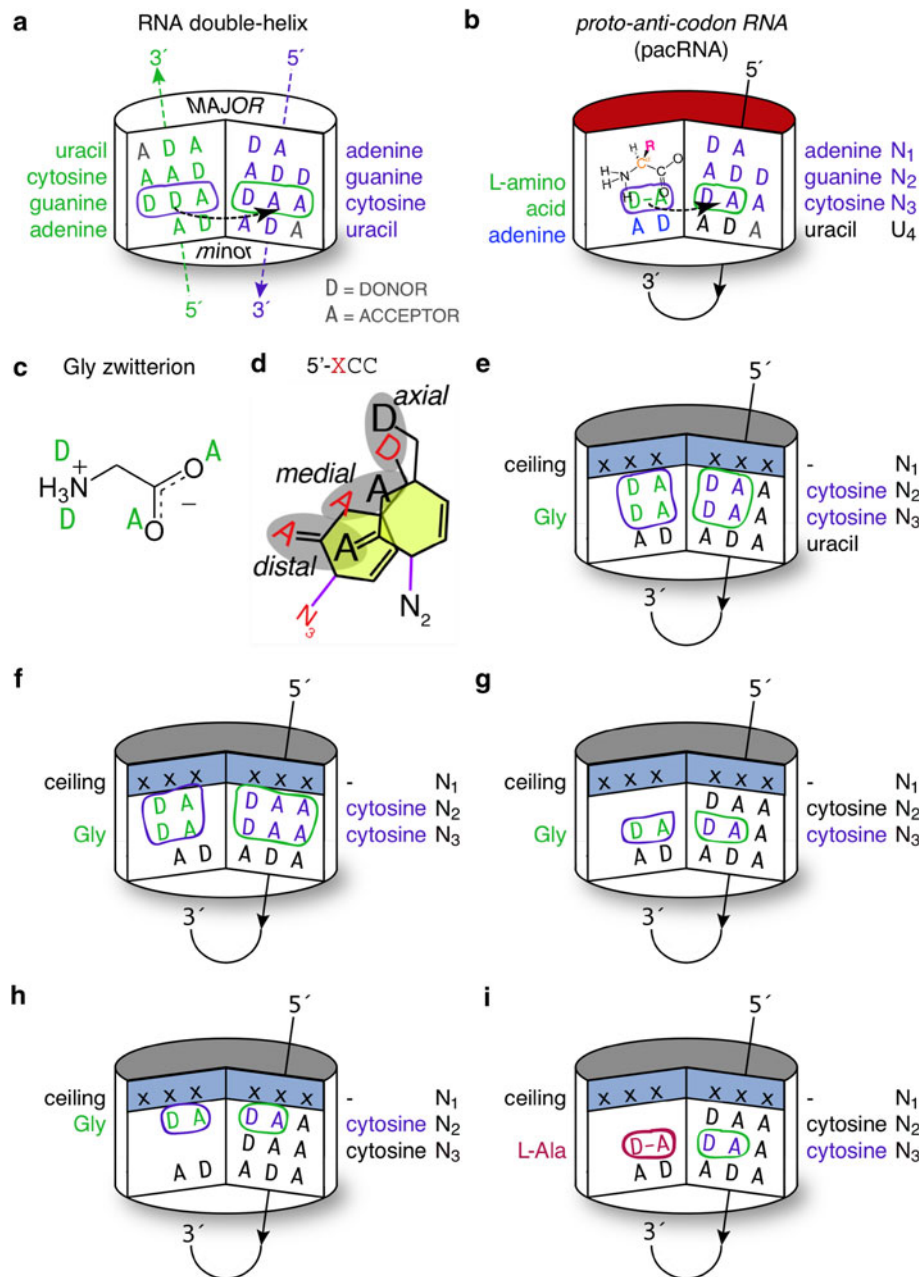
Fig. 3 The anti-codon pocket 5′-CC is specific for Gly. **a** Side view of a cross-section of a double-stranded RNA helix (*cylinder*) showing all four possible nucleotide base-pairs and their complementary Watson–Crick edges. The two anti-parallel RNA strands are shown in *green* and *purple*. The hydrogen-bond donor (D) and acceptor (A) profiles of each nucleotide are positioned along a row within each tetra-nucleotide surface and occur over the axial, medial, and distal columns along the cylindrical radius. Complementary groups on either side line up when the two surfaces are swung together (*green curved arrow*) while keeping the helical center in a fixed position. **c** Skeletal diagrams of the glycine (Gly) molecule and its physiological zwitterionic form. **d** The axial, medial, and distal A/D atoms of the Watson–Crick edges of the di-cytosine anti-codon pocket are depicted in red for N_3 and black for N_2. **e–g** Depicted are Gly-binding

configurations that are correctly oriented for charging to an adenosine-bearing molecule that is base-paired with uracil 3′ of the anti-codon sequence, 5′-$N_1N_2N_3$. The donor (D) and acceptor (A) groups for Gly and the anti-codon dinucleotide are shown in *green* and *purple*, respectively. In the proposed model, an overhanging "roof" (*blue shaded* nucleotides), which is provided by the first base-pair in the stem duplex functions as a ceiling over the Gly-binding pocket to prevent entry by most amino acids. **h** Additional Gly-binding states are available which are not in position for charging but which are related to the chargeable binding ensemble by simple rotations or translations. **i** The next best potential ligand for this surface is alanine (Ala, fuschia), which unlike Gly, possesses a limited ensemble of binding configurations due to its additional methyl group

Anti-Codon 5′-$X_1G_2G_3$ and L-Pro

Of all nine potential slots on all anti-codon surfaces, only the axial and medial slots of N_3 can coordinate the carboxy terminal end of the amino acid in the chargeable position. In this context, I find that the anti-codon 5′-$X_1G_2G_3$ is particularly suitable for an inverted ligand coordination that is appropriate to the puckering of the constrained three-carbon aliphatic side ring of L-proline (L-Pro). I propose that this anti-codon sequence binds L-Pro stably with a two binding state ensemble characterized by a carbonyl oxygen H-bond at the medial position of N_3 (Fig. 4c). One of these is coordinated by four H-bonds, but leaves the carboxyl-end carbon protected by a carbonyl oxygen and therefore protected from nucleophilic attack by the 3′-OH group of the adenosine nucleotide. However, this four H-bond coordination could be rolled over into the two H-bond pattern,

thereby exposing the carboxyl-end carbon in the inferred charging position (Fig. 4c). D-proline (D-Pro) might be able to bind this anti-codon sequence as well as L-Pro. However, I find that D-Pro's inverted puckering would be a disadvantage for subsequent charging because it raises the carboxyl carbon above the side chain ring and away from adenosine's 3′-ribose oxygen. The compact and constrained nature of the L-Pro side-chain allows it to be stably bound by an entry ensemble of four states able to roll-over the guanines into the chargeable ensemble (Fig. 4d).

Anti-Codon 5′-$X_1A_2C_3$ and L-Val

The cytosine in the third position of 5′-$X_1A_2C_3$ orients L-Val as it does for L-Gly and L-Ala anti-codons (Fig. 4e). However, the adenine base in the second position is nicely dovetailed to the aliphatic terminal "V" shape of valine.

Fig. 4 The anti-codon pockets for short-chained L-amino acids. **a** After Gly, the next five amino acids with the shortest side chains are shown here with their anti-codons. Correct charging of these six amino acids by these dinucleotide sequences requires a steric ceiling base-pair (*blue shading*) to preclude entry by longer chained amino acids. **b** The chargeable binding position for L-Ala. **c** and **d** The chargeable binding position for L-Pro (**c**) and entry binding configurations (**d**) related by a simple rolling movement (*curved arrow*) into the chargeable position. **e–g** The chargeable binding positions for L-Val, L-Thr, and L-Ser
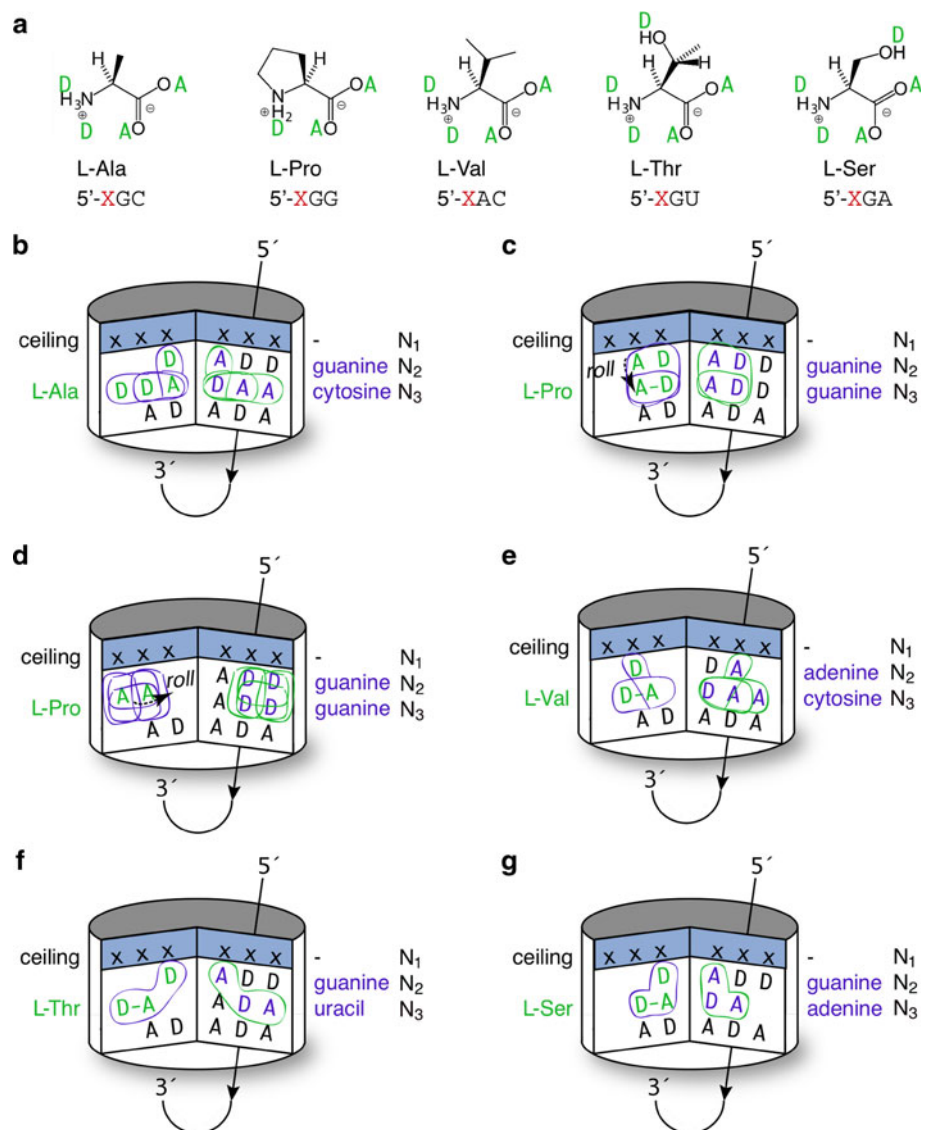
Compared to the Ala anti-codon (Fig. 4b), the hydrogen and amine groups are missing in just the right amount to accommodate valine.

Anti-Codons $5'-X_1G_2U_3$ and $5'-X_1G_2A_3$, and L-Thr and L-Ser

The anti-codon motif $5'-X_1G_2W_3$ may use the helically deep acceptor oxygen on the N_2 guanine to form a hydrogen-bond with the –OH hydrogen of both L-Thr (Fig. 4f) and L-Ser (Fig. 4g). I also further propose that the complementary use of either N_3 uracil or N_3 adenine further serves to distinguish between both ligands for two reasons. First, the alternate Watson–Crick profiles of uracil and adenine fix the chargeable binding positions at different radial slots productively. Second the compact N_3 uracil accommodates the extra methyl group of Thr, while the N_3 adenine prohibits accommodation of L-Thr.

Anti-Codons $5'-R_1C_2U_3$ and $5'-R_1C_2A_3$, and L-Ser and L-Cys

The anti-codon $5'-R_1C_2U_3$ may be bound by L-Ser at medial and axial N_3 columns if it is in an inverted orientation, suggesting an origin for the alternate L-Ser anti-codon (Fig. 5a, b). In addition, to this component of its chargeable binding ensemble, two additional chargeable binding states exist at the medial and distal columns (Fig. 5c). These are characterized by three H-bonds.

The related amino acid L-Cys replaces L-Ser's side chain oxygen with sulfur, which like oxygen possesses six electrons in the outermost energy levels. However, sulfur is a larger atom than oxygen, in addition to having a longer $C\beta$-SH bond compared to the $C\beta$-OH bond (Fig. 5a). Correspondingly, the predicted ligand binding pocket for L-Cys replaces U_3 for A_3, while keeping the R_1C_2 sequence of the alternate Ser anti-codon (Fig. 5d). This $5'-R_1C_2A_3$ sequence necessitates using the axial and medial columns for N_3 coordination, increases the distance between N_3 and N_2 coordinating groups and precludes coordination of L-Ser.

Anti-Codon $5'-X_1A_2R_3$ and $5'-X_1A_2U_3$, and L-Leu, L-Ile

I propose that the pacRNA anti-codon sequence $5'-X_1A_2R_3$ evolved into the two modern L-Leu anti-codons $5'$-NAG and $5'$-YAA later during the evolution of proto-tRNAs and translation (Fig. 5e, f). I note three chargeable binding configurations for this sequence, all of which allow entry and side chain packing as a result of an empty distal column of the N_2 adenine. These binding states are related either by a 90° forward or 90° lateral rotation (see arrows in Fig. 5f).

The anti-codon sequence $5'-Y_1A_2A_3$ is similar to the $5'-X_1G_2G_3$ anti-codon for L-Leu in also having two purines

at N_2 and N_3 (Fig. 5g). However, by having adenine instead of guanine at N_3, the ligand must necessarily use the helically deep D atom of adenine to coordinate the carboxyl terminal unlike in $5'-X_1A_2G_3$. This difference is actually a minor one given the angle of the Watson–Crick edge of adenine at N_3 and the degree of rotational freedom in the $C\alpha$–$C\beta$ bond.

The related anti-codon $5'-X_1A_2U_3$ resembles the L-Leu anti-codon motif but replaces the N_3 adenine with a more compact pyrimidine, uracil, which requires coordination at the medial and distal columns of N_3 (Fig. 5h). For this anti-codon, I note three binding configurations, related by clock-wise rotation, for its cognate L-isoleucine (L-Ile). These rotations maintain the H-bond at the medial N_3 column by the carbonyl oxygen on the amino acid, while accommodating the methyl-group at the first side-chain methylene within adenine's empty distal pocket. I suggest that the preference for any nucleotide except cytosine at N_1 may have evolved later during tRNA evolution to avoid mis-specification with the L-Met anti-codon.

pacRNAs for Amide or Acidic R-Groups

I find that the cognate anti-codons for the L-Asn, L-Asp, L-Gln, and L-Glu can distinguish the following features when provided with pacRNA $5'$-ceilings as indicated (Fig. 6). The amide side chains of asparagine (Asp) and glutamine (Gln) do not ionize, but they each provide polarized hydrogen-bond donors and acceptors (Fig. 6a) (Creighton 1993). In contrast, the carboxyl groups of aspartate (Asp) and glutamate (Glu) do ionize under physiological conditions, and these side chains provide two hydrogen-bond acceptors (Fig. 6a) (Creighton 1993). The other significant difference among these amino acids is that Gln and Glu possess side chains that are one methylene group longer than those on Asn and Asp (Fig. 6a).

Either of the pyrimidine bases at N_1 suffices for L-Gln and L-Glu because each provides A and D atoms at axial and medial columns. Furthermore, the receded Watson–Crick edges of these bases are provided at the right distances. In contrast, I propose that the L-Asn and L-Asp pacRNA anti-codons had no preference for any nucleotide at N_1 because I infer this position to have been in the stem duplex. Instead, I propose that the use of N_1 pyrimidine bases in L-Gln, L-Lys, and L-Glu pacRNAs predetermined the use of N_1 purines in future tRNA anti-codons for L-His, L-Asn, and L-Asp (Fig. 6b–e).

Another important feature is the use of uracil or guanine at N_3 by the amino acids with amide side chains, versus cytosine at N_3 by the amino acids with carboxyl group side chains. This difference fixes the key carbonyl oxygen H-bond at the medial position for L-Asn and L-Gln (Fig. 6b, d), and the axial column for L-Asp and L-Glu (Fig. 6c, e).
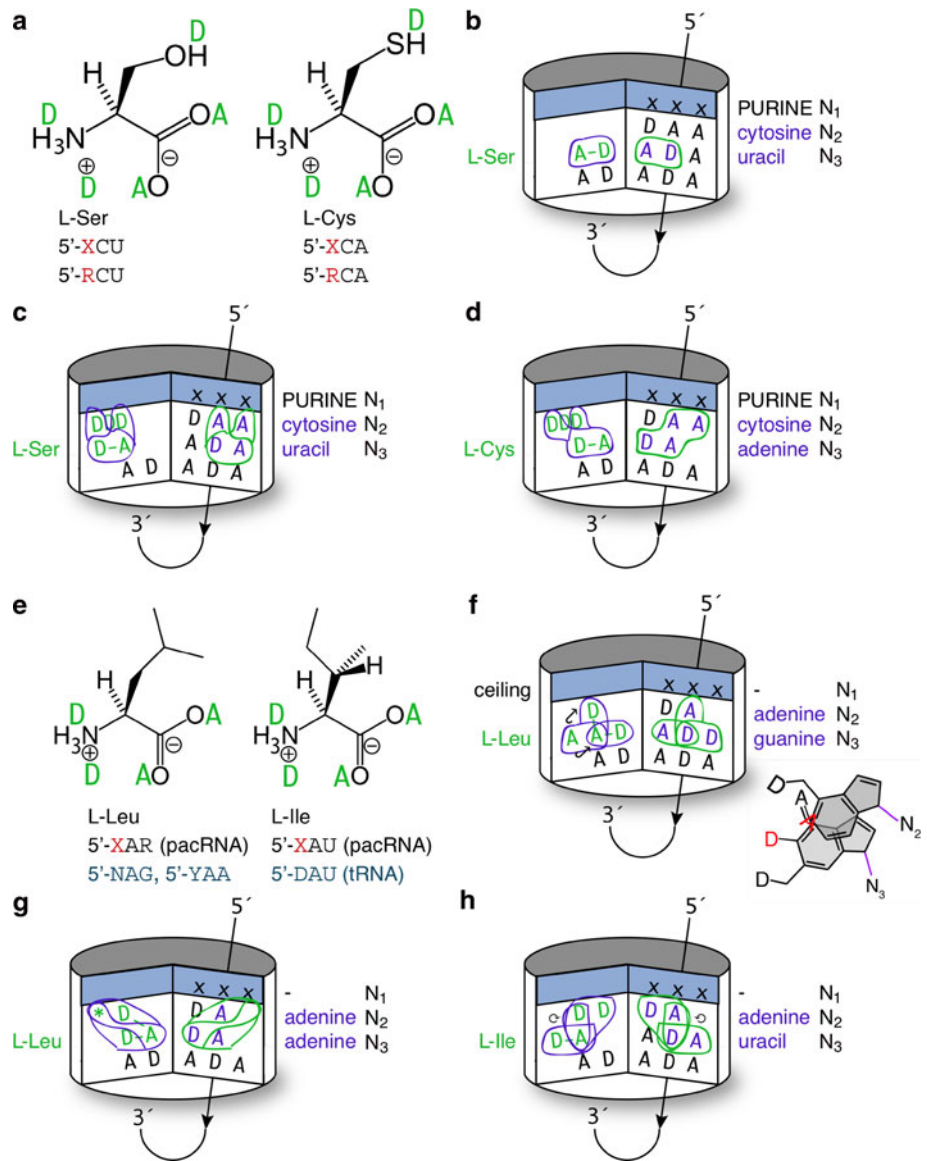
This difference thus reduces the number of chargeable positions for their cross-antagonists.

pacRNAs for Long-Chained R-Groups

Anti-Codon 5′-$C_1A_2U_3$ and L-Met

The axial and medial columns within the sequence 5′-$C_1A_2U_3$ are devoid of any H-bond acceptor oxygens and their electronegativity. It is also entirely devoid of any acceptor atoms in the axial columns. I propose that these features predisposed this specific anti-codon sequence to the L-Met side chain, which is very long, non-polar, and unreactive (Creighton 1993) (Fig. 7a, b). The 5′-CA sequence, which binds and charges cysteine when it occurs at N_2 and N_3, would also bind when it occurs at N_1 and N_2

as it does in the 5′-CAT anti-codon sequence. However, Cys bound at this position would be unchargeable because it would be elevated one nucleotide too high for aminoacylation. Thus, Cys would pose only as a reversible binding competitor of this binding pocket.

Anti-Codon 5′-$X_1U_2G_3$ and L-His

The imidazole side chain of His is usually protonated up to physiological pH, and the extra positive charge is shared by the nitrogen atoms via resonance (Creighton 1993). It consequently has two donor groups for hydrogen bonds. I therefore propose that the 5′-$X_1U_2G_3$ anti-codon coordinated the ionized form of L-His and that its entry into this pocket was side-ways, an unusual distinction from most other amino acids (Fig. 7c). In this orientation, the carboxyl

Fig. 6 Anti-codon pockets for L-amino acids with amide or acidic groups. **a–e** The amide-side and carbonyl groups of L-asparagine (L-Asn), aspartate (L-Asp), L-glutamine (L-Gln), and L-glutamate (L-Glu) result in exclusive chargeable binding ensembles concentrated in the outlined surfaces as shown
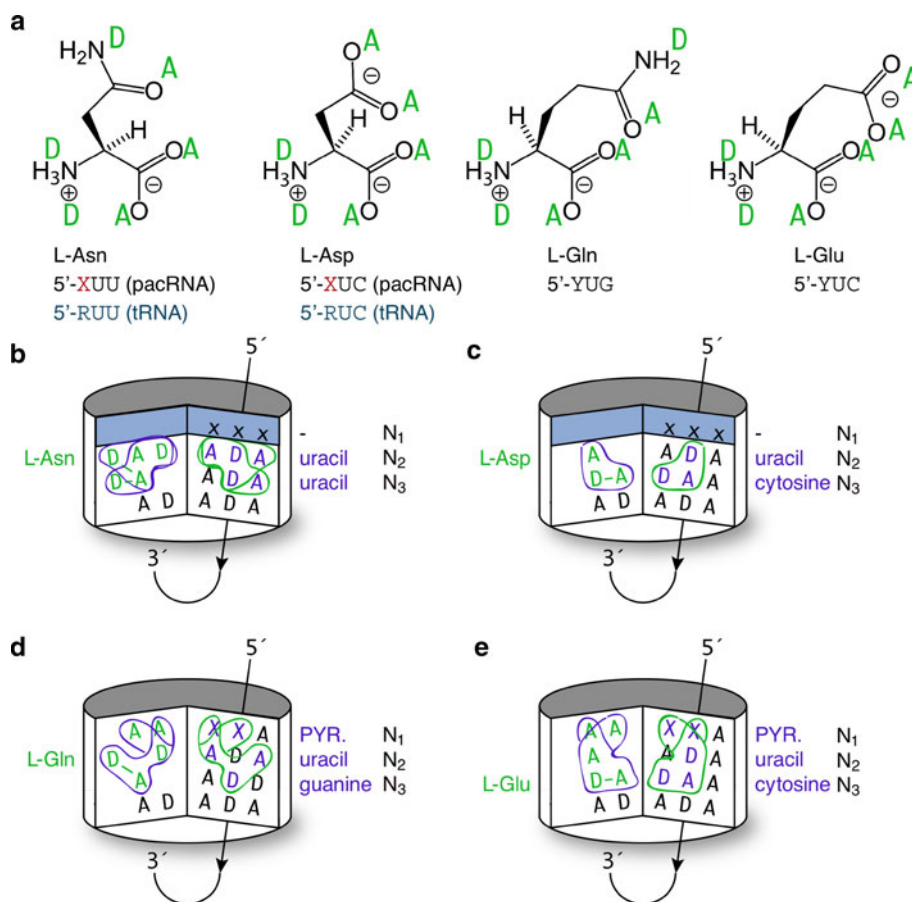
end is correctly oriented and the imidazole nitrogen donor groups make hydrogen bonds with the helically deep acceptor groups of uracil and guanine. There is thus only a single chargeable binding configuration for L-His. Similar to our argument for the amide side-chains, the use of a purine at N_1 may be necessary to prevent anti-codon access to L-Gln.

Anti-Codon $5'$-$Y_1U_2U_3$ and L-Lys

The long chain of L-Lys ends with an amino-group and a hydrogen bond donor (Fig. 7a). Its cognate sequence $5'$-$Y_1U_2U_3$ fills the entire pocket with acceptor groups along the helically deep column and the along N_1 (Fig. 7d). N_1 is a pyrimidine, which will have an A atom at the distal Watson–Crick edge and a second one in the medial or axial positions. I therefore propose that this coordinated L-Lys across the diagonal.

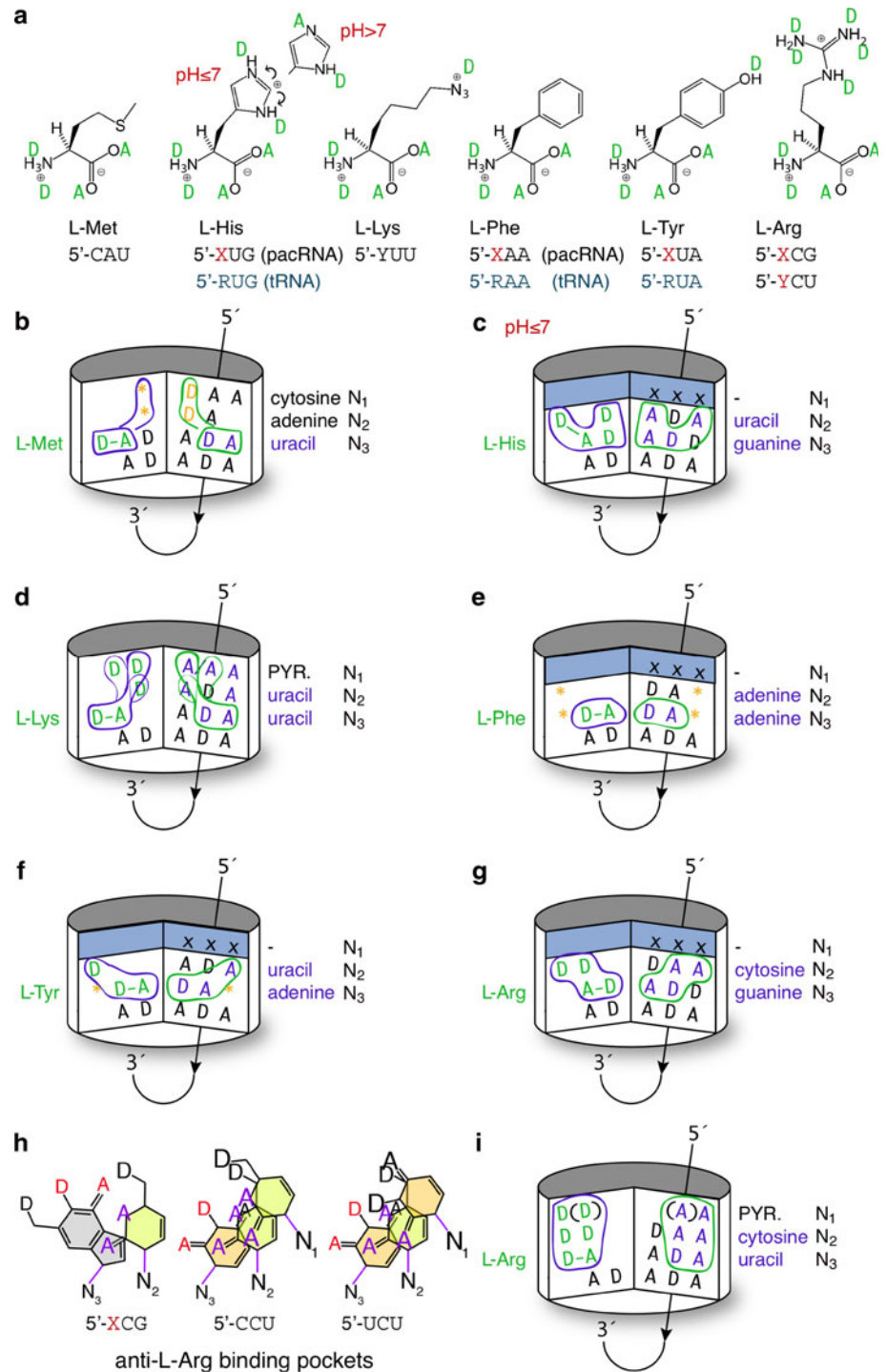*Anti-Codons $5'$-$R_1A_2A_3$ and $5'$-$R_1U_2A_3$, and L-Phe and L-Tyr*

I note one unconstrained binding configuration for L-Phe with its cognate anti-codon (Fig. 7e). Notably, it uses

double adenines at N_2 and N_3 as these are devoid of A/D atoms at the distal columns and facilitates entry and chargeable packing of the phenolic ring side chain. By replacing the N_2 adenine with uracil, an acceptor atom is provided at the distal column that can accept a hydrogen bond from L-Tyr's extra hydroxyl group (Fig. 7f).

Anti-codons $5'$-$X_1C_2G_3$ and $5'$-$Y_1C_2U_3$ and L-Arg

Arginine (Arg) is an unusually long amino acid with a strongly basic δ-guanido group, which is ionized over a wide pH range (Creighton 1993) (Fig. 7a). Consequently, L-Arg is used in a structural capacity in protein folding for its ability to participate in up to five hydrogen bonds via the side chain alone (Borders et al. 1994). Resonance charge transfer in its guanido group gives it a planar character (Creighton 1993), which I propose facilitates pocket entry and packing of the L-Arg side chain. Based on the number of H-bonds in the chargeable configuration, the following anti-codon sequences would have had increasingly higher affinities for L-Arg: $5'$-$X_1C_2G_3$ (four H-bonds) < $5'$-$U_1C_2U_3$ (five H-bonds) < $5'$-$C_1C_2G_3$ (six H-bonds). These sequences correspond to L-Arg cognate anti-codons in the extant codon table (Fig. 7g–i).

The $5′-X_1C_2G_3$ and $5′-Y_1C_2U_3$ motifs coordinate the amino and carboxy termini of L-arginine (L-Arg) in two anti-parallel orientations that share in common only the placement of the carboxy-terminus of L-Arg at the medial column (Fig. 7a, g–i). The $5′-X_1C_2G_3$ would coordinate L-Arg in an unusual inverted position with the amino group at the axial position (Fig. 7g, h) similar to L-Pro. The uniquely long side chain of L-Arg, which has six atoms/bonds past Cα, is able to bind in this inverted position. It would be able to reach around and make H-bonds with medial and distal A atoms on the N_2 cytosine but be unable to reach N_1, which I infer is a paired ceiling.

By replacing the N_3 guanine with uracil in the motif $5′-Y_1C_2U_3$, L-Arg can be coordinated by the medial/distal A/D

atoms of the N_3 Watson–Crick edge. As its side chain no longer needs to wrap around as it did for the 5′-$X_1C_2G_3$ pacRNA pocket, L-Arg is now able to make H-bonds at both N_1 and N_2 provided that they are both pyrimidines (Fig. 7h, i).

pacRNAs with Prohibitive Pockets

When 5′-N_1N_2 consists of a receded pyrimidine wall, followed by an adenine at N_3, there is an unusually long distance separating the N_3 Watson–Crick edge from those on N_1N_2 (Fig. 8a–d, blue planes). Furthermore, of all the nucleotide bases, adenine is alone in having only two groups for forming hydrogen bonds at its Watson–Crick edge. Consequently, the 5′-$Y_1Y_2A_3$ motif is not well-suited for coordinating an amino acid in the pacRNA-mediated aminoacylation system. These sequences would coordinate only extremely long and bulky side-chained amino acids that can be stabilized across both sides of the gap prior to charging. Therefore, it is an interesting confirmation of the model to find that the sequences fitting this ill-suited motif are associated with all three stop codon sequences: anti-ochre (5′-UUA, Fig. 8a), anti-amber (5′-CUA, Fig. 8b), and anti-opal (5′-UCA, Fig. 8c). This anti-codon pattern is also found in the anti-codon 5′-CCA (Fig. 8d), which has been assigned to the bulkiest natural amino acid, L-tryptophan (L-Trp) (Fig. 8e). If the N_3 adenine base is replaced with guanine, the angled Watson–Crick profile of guanine

Fig. 8 Termination codons are predetermined by anti-codon stereochemistry. **a–d** *Top–down* view of anti-codon sequences matching 5′-$Y_1Y_2A_3$ looking down the axis of a helical turn. These sequences have unusually large gaps (*blue* planar surfaces) between the D/A atoms (*red letters*) of the purine adenine and the D/A atoms of the pyrimidine dincleotide. Nucleotide bases stack downward into the page, and glycosidic bonds to the ribose C1′ atom are shown in *purple*. These anti-codons are complements of all three translational termination signals, **a** ochre (5′-UAA), **b** amber (5′-UAG), and **c** opal (5′-UGA); and **d** the codon 5′-UGG for the bulkiest amino acid, L-Tryptophan (L-Trp). **e** The L-Trp zwitterion with its D and A atoms is shown in an orientation that would fit into the anti-codon sequence in **d** if the side chain was tilted downwards into the page. **f** The anti-L-Gln sequence 5′-UUG is shown for comparison. This sequence has the purine guanine instead of an adenine at N_3 without creating a large gap (*blue* planes) to the receded pyrimidine bases at N_2 and N_1

does not produce much of a gap between it and the receded pyrimidine wall (Fig. 8f).

Discussion

I showed that pacRNA anti-codon sequences naturally coordinate their cognate L-amino acid ligands for charging to a specific adenosine nucleotide that is base-paired immediately 3′ of the anti-codon sequence. This fixed coordinated position would have allowed the pacRNA molecule to activate the amino acid and charge it to its 3′-end. This stereochemical RNA receptor model complements recent advances from ribozyme biochemistry. For example, recent reviews of RNA selection experiments for amino acid binding found a general correlation for cognate anti-codon triplets amino acid-binding RNA sequences (Yarus et al. 2009; Rodin et al. 2011). These results provide support for a stereochemical RNA era. The proposed pacRNA model adds to these perspectives by explaining evolutionary origins of homochirality, stop codons, and third nucleotide synonymous positions of codons. The pacRNA model also introduces another aspect to nucleotide base complementarity as fundamental as base pairing (Watson and Crick 1953).

One intriguing explanation for homochirality has been that potential sources of circularly polarized starlight in the early universe could have selectively degraded D-amino acids (Bailey et al. 1998). However, it is not clear how this effect could have been absolute (Irion 1998). For example, "twisted starlight" would have to prevent the assignment of all codons to any D-amino acid. The pacRNA model implies instead that amino acid homochirality is intrinsic to its stereochemical relationship with nucleic acids. Nonetheless, the pacRNA model assumes D-ribose chirality. So an alternative explanation is that "twisted starlight" might have biased the chirality of nucleic acid sugars indirectly through its relationship with amino acids. This latter explanation would not require complete dominance of L-amino acids over D-amino acids in order for D-ribose-based nucleic acid systems to win over L-ribose nucleic acid systems. Once D-ribose nucleic acids had taken over, they would have intrinsically selected the use of their reciprocal counterparts, the L-amino acids. Thus, the pacRNA model appears to fill an important gap in the original proposal connecting "twisted starlight" to amino acid homochirality.
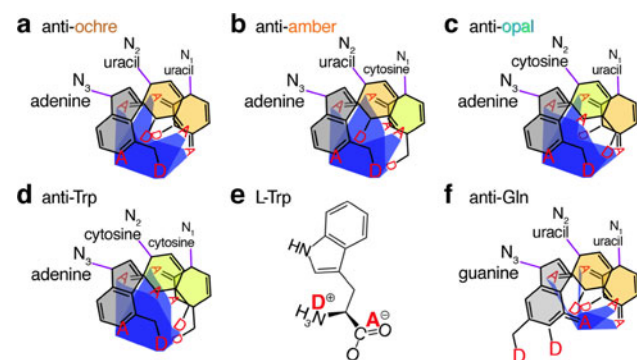
An Early Aminoacylated RNA World

The proposed *pacRNA* model implies a natural order of events for evolutionary transitions in an early RNA or RNA-like world (Gilbert 1986; Bartel and Unrau 1999; Schöning et al. 2000). The intrinsic amino acid coordination capacities of pacRNAs suggest that the beginnings of RNA world are

already an aminoacylated RNA world. As such, aminoacylated RNA world would be more capable of constructing an RNA-based metabolism and genetic system than currently envisioned for RNA world (Gilbert 1986).

Aminoacylated RNA world would have been characterized by the natural, evolutionary progression through three stages featuring additional functional groups: (I) pacRNAs (II) *cis*-element codons, and (III) non-ribosomal and/or pre-ribosomal peptide polymerization. I suggest that the capacity to form free, aminoacylated, oligo-ribonucleotides with long-chained hydrophobic amino side chains may have been selectively maintained for their immediate use in rudimentary lipid-like membranes and/or hydrophobic anchoring pockets ("aminoacylated RNA world I"). Once aminoacylated a pacRNA could have used its exposed anti-codon pocket for *cis–trans* interactions. Subsequent evolution of codon *cis*-elements on ribozymes may have allowed the formation of hydrophobic patches that assisted in ribozyme folding ("early aminoacylated RNA world II"). Codons could also have evolved to recruit amino acid chemical groups for their use in enzymatic catalysis, thus broadening use of the amino acid repertoire. Such codon/anti-codon linkages in this RNA-world may have rapidly increased in density over the surfaces of most RNA gene products, leading to an era of complex amino acid-ribozyme intermediates ("advanced aminoacylated RNA world II"). Continued evolutionary increases in the density of L-amino acids coating all ribozymes would have led naturally to their catalyzed polymerization on ribozyme surfaces ("aminoacylated RNA world III"). Several biochemical polypeptide molecules in extant organisms have unusual linkages not seen in proteins; their use may have originated from these latter stages of a mature aminoacylated RNA world. One example is the cellular redox agent, L-glutathione (γ-L-Glutamyl-L-cysteinylglycine), which contains a peptide linkage between the amine group of cysteine and the side-chain carboxyl group of glutamate.

Aminoacylated RNA world would have come to an end with the overwhelming success of organisms in the stem LUCAn lineage. This lineage would have been characterized by an increasingly specialized ribosome-mediated peptide polymerization system and deprecated usage of pacRNA-cofactor-dependent ribozymes. In the multi-stage model of aminoacylated RNA world predicted by the pacRNA model, the first proto-tRNAs would have been evolutionary exaptations of pacRNAs in the stem LUCAn ancestors.

Materials and Methods

Modeling of RNA Ligand Coordination

I verified core aspects of the pacRNA model using various modeling media, including physical 3-dimensional chemical bonding models, a custom-made multi-planar transparency viewer, and solved RNA-structure-based computational assembly (Assemble ver. 1.0) of anti-codon trinucleotides (Jossinet et al. 2010). These were visualized using the Assemble software or with PyMol in order to visualize superimposed amino acid molecules. Because the exposed single-stranded anti-codon pocket may "breath" and is more flexible than the double-stranded stem structures, I made an allowance of approximately $\pm 1/2$ of the corresponding hydrogen bond length to score potential H-bond configurations. I then annotated each such bonding potential on the cylindrical graphical representation, and attempted to show all relevant binding configurations that would be considered for the chargeable binding ensemble, i.e., binding configurations related by a simple rotation or translation into the charging configuration. I also modeled all other potential competitors and inferred the presence of a 5′ base-paired nucleotides (steric ceiling) when it was necessary to regain specificity. I only considered 3′-aminoacylation because the 2′-OH on the ribose points away from the anti-codon surface while the 3′-OH group points toward it. Nonetheless, I cannot formally exclude the possibility that some amino acid ligands were charged to the 2′ group to some extent.

Modeling of Energetics

I considered that the energetics of charging is effected by several important variables, which may have differed for different pacRNAs at different times. First, in addition to the primary catalytic advantage provided by correct spatial coordination of the ligand by the anti-codon pocket, various pacRNA systems could have used additional oligonucleotide catalytic mechanisms present within the pacRNA molecule, the acceptor stem molecule if this was not already part of the pacRNA, or other molecules in *trans*. However, I allow the possibility that this may not have been necessary. Second, the energetics are necessarily dependent on local concentrations of ligands and of the individual pacRNAs for each ligand. Third, the local concentration of ions within the solution must necessarily be assumed. I therefore consider that the global correspondence between the extant codon table and the natural chargeable binding configurations in pacRNAs impels us to use the model to constrain the ancestral parameters describing the concentration ranges for ligands, antagonists, and ions. Here I focused on presenting the H-bond binding ensembles, which will be necessary for this future use of the model.

References

Bailey J, Chrysostomou A, Hough JH, Gledhill TM et al (1998) Circular polarization in star-formation regions: implications for biomolecular homochirality. Science 31:672–674

Bartel DP, Unrau PJ (1999) Constructing an RNA world. Trends Cell Biol 9:M9–M13

Borders CL Jr, Broadwater JA, Bekeny PA, Salmon JE et al (1994) A structural role for arginine in proteins: multiple hydrogen bonds to backbone carbonyl oxygens. Protein Sci 3:541–548

Creighton TE (1993) Proteins: structures and molecular properties, 2nd edn. W. H. Freeman and Company, New York

Deutscher MP (1982) tRNA nucleotidyltransferase. The Enzymes 15:183

Di Giulio M (2006) The non-monophyletic origin of the tRNA molecule and the origin of genes only after the evolutionary stage of the last universal common ancestor (LUCA). J Theor Biol 240:343–352

Fujishima K, Sugahara J, Tomita M, Kanai A (2008) Sequence evidence in the archaeal genomes that tRNAs emerged through the combination of ancestral genes as 5′ and 3′ tRNA halves. PLoS ONE 3:e1622

Gilbert W (1986) The RNA world. Nature 319:618

Hopfield JJ (1978) Origin of the genetic code: a testable hypothesis based on tRNA structure, sequence and kinetic proofreading. Proc Natl Acad Sci USA 75:4334–4338

Irion R (1998) Did twisty starlight set stage for life. Science 31:626–627

Jossinet F, Ludwig TE, Westhof E (2010) Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. Bioinformatics 26:2057–2059

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

Maizels N, Weiner AM (1994) Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. Proc Natl Acad Sci USA 91: 6729–6734

Marck C, Grosjean H (2002) tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. RNA 8:1189–1232

Pan B, Xiong Y, Steitz TA (2010) How the CCA-adding enzyme selects adenosine over cytosine at position 76 of tRNA. Science 330:937–940

Randau L, Munch R, Hohn MJ, Jahn D, Soll D (2005) *Nanoarchaeum equitans* creates functional tRNAs from separate genes for their 5′- and 3′-halves. Nature 433:537–541

Rodin AS, Szathmáry E, Rodin SN (2011) On origin of genetic code and tRNA before translation. Biol Direct 6:14

Schöning KU, Scholz P, Guntha S, Wu X, Krishnamurthy R, Eschenmoser A (2000) Chemical etiology of nucleic acid structure: the α-threofuranosyl-(3′→2′) oligonucleotide system. Science 290:1347–1351

Sugahara J, Yachie N, Sekine Y, Soma A, Matsui M et al (2006) SPLITS: a new program for predicting split and intron-containing tRNA genes at the genome level. In Silico Biol 6:411–418

Sugahara J, Yachie N, Arakawa K, Tomita M (2007) In silico screening of archaeal tRNA-encoding genes having multiple introns with bulge-helix-bulge splicing motifs. RNA 13:671–681

Watson JD, Crick FH (1953) A structure for deoxyribose nucleic acid. Nature 171:737–738

Weiner AM, Maizels N (1999) The genomic tag hypothesis: modern viruses as molecular fossils of ancient strategies for genomic replication, and clues regarding the origin of protein synthesis. Biol Bull 196:327–328 discussion 329–330

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

Yarus M, Widmann JJ, Knight R (2009) RNA-amino acid binding: a stereochemical era for the genetic code. J Mol Evol 69:406–429