tRNA acceptor stem and anticodon bases form independent codes related to protein folding

Charles W. Carter Jr.[1] and Richard Wolfenden[1]

Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599-7260

Aminoacyl-tRNA synthetases recognize tRNA anticodon and 3′ acceptor stem bases. Synthetase Urzymes acylate cognate tRNAs even without anticodon-binding domains, in keeping with the possibility that acceptor stem recognition preceded anticodon recognition. Representing tRNA identity elements with two bits per base, we show that the anticodon encodes the hydrophobicity of each amino acid side-chain as represented by its water-to-cyclohexane distribution coefficient, and this relationship holds true over the entire temperature range of liquid water. The acceptor stem codes preferentially for the surface area or size of each side-chain, as represented by its vapor-to-cyclohexane distribution coefficient. These orthogonal experimental properties are both necessary to account satisfactorily for the exposed surface area of amino acids in folded proteins. Moreover, the acceptor stem codes correctly for β-branched and carboxylic acid side-chains, whereas the anticodon codes for a wider range of such properties, but not for size or β-branching. These and other results suggest that genetic coding of 3D protein structures evolved in distinct stages, based initially on the size of the amino acid and later on its compatibility with globular folding in water.

genetic code | aminoacyl-tRNA synthetases | urzymes | multivariate regression | protein folding

The genetic code is implemented by two distinct superfamilies of protein–RNA complexes between an aminoacyl-tRNA synthetase (aaRS) from one of two classes (1, 2) and its cognate tRNA. These recognition complexes effect the transfer of activated amino acids to the correct tRNA molecule, producing aminoacyl-tRNAs needed for protein synthesis by the ribosome. Errors in charging are rare (3, 4), and it is generally agreed that the low frequency of mischarging is based on synthetase recognition of specific identity elements in tRNA molecules (5). Many investigators (6–8) have observed that the codon table tends to reduce deleterious effects of point mutations (9) by assuring that they do minimal violence to the physical requirements of protein folding. One earlier study (10) identified a nonrandom tendency for hydrophilic side-chains to be coded by an A as the second codon base, hinting at more extensive relationships between the code and factors that direct protein folding.

tRNA identity elements (5, 11) map to both the anticodon and acceptor stem at opposite ends of the L-shaped tRNA molecule and are distinct from binding determinants for elongation factor-Tu in the T-stem (12). Invariant cores of both classes of aaRS, termed urzymes (from the prefix ur- = primitive), lack anticodon-binding domains and cannot recognize the anticodon. However, they catalyze amino acid activation and acyl transfer with K_M values approaching those of contemporary aaRS, consistent with their participation in early protein synthesis (13–16). The implied ability of ancestral aaRS to recognize tRNA acceptor stems, but not anticodons, is consistent with the suggestion that the earliest proteins were coded not by anticodon–codon interactions but by a more plastic "operational RNA code" exclusively in the accepter-stem (17, 18), If an acceptor stem code could be identified, it might furnish clues about what features of the earliest coded peptides gave them a selective advantage. Here, we identify separate anticodon and acceptor stem codes and the amino acid properties that determine them.

Results

Amino Acid Side-Chain Sizes and Polarities (10). The folding of a protein is believed to depend on interactions of its constituent amino acids with each other and with solvent water. Both types of interaction can be modeled experimentally by determining equilibria of transfer of amino acid side mimics (with propionamide, for example, representing glutamine) among aqueous solution, the nonpolar solvent cyclohexane, and the vapor phase (19) (SI Appendix, Fig. S1). These transfer equilibria (and free energies, $\Delta G = -RT\ln K_{eq}$) can be considered to measure the principal forces stabilizing protein structures as follows.

Vapor-to-water transfer equilibria ($K_{v>w}$, where v = vapor and w = water) describe a molecule's hydrophilic character (20), i.e., its absolute tendency to leave the vapor phase, in which each solute molecule exists in isolation at ordinary temperatures and pressures and enter water at infinite dilution. Water-to-cyclohexane transfer equilibria ($K_{w>c}$, where c = cyclohexane) furnish a measure of what is often termed hydrophobicity (21), i.e., a molecule's tendency to leave water and enter a nonpolar condensed phase. Vapor-to-cyclohexane transfer equilibria ($K_{v>c} = K_{v>w} \times K_{w>c}$) measure the van der Waals forces that attract a solute from the vapor phase to the walls of a nonpolar solvent cavity minus the cost of making that cavity (SI Appendix, section 1.). Empirically, $\Delta G_{v>c}$ values have been found to be closely related to a molecule's size or accessible surface area (19), defined as the area over which the center of a water molecule can retain van der Waals

Significance

The universal genetic code is the earliest point to which we can trace biological inheritance. Earlier work hinted at a relationship between the codon bases and the physical properties of the 20 amino acids that dictate the 3D conformations of proteins in solution. Here, we show that acceptor stems and anticodons, which are at opposite ends of the tRNA molecule, code, respectively, for size and polarity. These two distinct properties of the amino acid side-chains jointly determine their preferred locations in folded proteins. The early appearance of an acceptor stem code based on size, β-branching, and carboxylate groups might have favored the appearance of antiparallel peptides that have been suggested to have a special affinity for RNA.

BIOPHYSICS AND COMPUTATIONAL BIOLOGY

CHEMISTRY

contacts with the side-chain in a gly-X-gly tripeptide without penetrating other atoms (22, 23).

Sizes, Polarities, and Protein Folding. Until now, participation of two aaRS classes in genetic coding has been rationalized as a result of successive binary choices (24) or as a means of avoiding coding ambiguity (25). Here, we show that this distinction appears to be related to the complementary roles of class I and class II amino acids in protein folding.

Members of subclass IA (Leu, Ile, Val, and Met) have aliphatic side-chains and are found in hydrophobic cores. Members of subclass IIA (Ser, Thr, and His) are small amino acids with water-favoring side-chains. Subclasses B (with carboxyl, amide, primary amine side-chains*) and C (aromatic) in both classes contain similar amino acids. Apart from those parallels, class I amino acids tend to be less strongly attracted to water (16) (Fig. 1A) and larger (26) (Fig. 1B) than those of class II. A more substantive (4σ) difference between class I and class II amino acids appears in the distributions of folded accessible surface area (ASA_{fold}) (27), which describes their situations after protein folding has implemented instructions conveyed by the genetic code. Class I amino acids tend to be buried (median ASA_{fold} = 0.32); those in class II remain largely on the surface (median ASA_{fold} = 0.54; Fig. 1C).

Amino acid core/surface distributions (log K_{surf}) (28) correlate only approximately with their $\Delta G_{w>c}$ ($R^2 = 0.53$). The correlation between observed and calculated distributions improves ($R^2 = 0.81$) if their $\Delta G_{v>c}$ is also taken into consideration as shown in Eq. 1

$$\log K_{Surf} = \beta_0 + (\beta_1 \times \Delta G_{w>c}) + (\beta_2 \times \Delta G_{v>c}) + \beta_3 \times (\Delta G_{w>c} \times \Delta G_{v>c}) + \varepsilon,$$ [1]

where coefficients β_{0-3} are estimated by least squares, and ε is an error term (Fig. 1D).[†] Thus, the behavior of an amino acid side-chain on folding can be predicted from its solvent transfer free energy (54%), the van der Waals contacts it can make in the folded state (18%), and their two-way interaction (9%; Fig. 1E). When two outliers, proline (whose presence in turns exposes its side-chain preferentially to solvent) and cysteine (whose participation in disulfide bonds and metal coordination tends to bury it) are omitted, R^2 increases from 0.81 to 0.91, with correspondingly low P values ($P_{\beta1} < 0.0001$, $P_{\beta2} < 0.0001$, and $P_{\beta3} = 0.0017$) (28).

Regression Models Uncover Different Codes in tRNA Anticodons and Acceptor-Stems. To relate transfer free energies to anticodon and acceptor stem bases, we tested regression models of the form in Eq. 2:

$$\Delta G_{w>c} \text{ or } \Delta G_{v>c} = \beta_0 + \sum \beta_i \times \text{Basebit}_i + \sum \beta_{ij} \times \text{Basebit}_i \times \text{Basebit}_j + \varepsilon,$$ [2]

where ε is a residual, β_i, $\beta_{ij...}$ are estimated coefficients, and {Basebit$_i$} are binary predictors, two for each identity element base, as illustrated schematically in Fig. 2A. The coding bits for the degenerate leucine anticodon are described in *Methods*. Optimal multivariate models are usually hard to identify unambiguously without exhaustive enumeration (29), but in this case, automated forward and backward stepwise searches led to the same optimum models. Most Student t test probabilities (*SI Appendix*, Tables S2 and S4) are <0.0001, indicating that the null hypothesis is highly unlikely to be true.

---

*Lysine is represented in Class IIB and also, to a far lesser extent, in IB.

[†]Substituting mass for $\Delta G_{v>c}$ leads to essentially the same result (*SI Appendix*, Fig. S2).
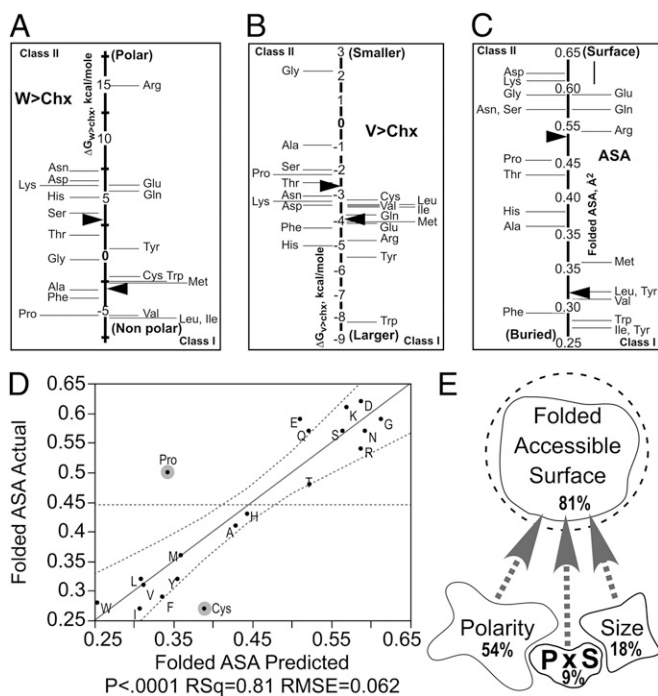
Fig. 1. Roles of amino acid size, polarity, and class in protein folding. (A) Distribution of $\Delta G_{w>c}$ (P = 0.5). (B) Distribution of $\Delta G_{v>c}$. Class I contains larger, less polar side-chains (P = 0.02). (C) Distribution of ASA_{Fold} for class I and II amino acids (ASA_{fold}; P = 0.03). (D) Bivariate regression model for ASA_{fold}. Size and polarity both determine final locations of amino acids. The linear combination of amino acid mass and $\Delta G_{w>c}$ yields essentially the same correlation (*SI Appendix*, Fig. S2). (E) Complementary effects of amino acid polarity (P < 0.0001) and size (P = 0.0002) account for most of the variance of ASA_{fold}. Relative contributions are indicated by the three irregular shapes. PxS, the two-way interaction, indicates the difference between the effect of polarity for smaller vs. larger amino acids. Plots prepared using JMP (63).

The many predictors necessary to discriminate between the 20 canonical amino acids leave the analysis vulnerable to overfitting and, in turn, to poor predictive performance. We surmounted this difficulty two ways. (i) We deleted coefficients that were small, relative to their uncertainty, and had little influence on the models, accepting only models with at least four more data points than parameters. (ii) We cross-validated models against predictions for two amino acids, selenocysteine (Sec) and pyrrolysine (Pyl), which lie outside the canonical training set but whose properties can be reliably estimated. Cross-validation (Fig. 2 D and E) discriminates decisively between the two tRNA coding regions.

The Anticodon and Acceptor-Stem Code Exactly for $\Delta G_{w>c}$ and $\Delta G_{v>c}$. Earlier (10), one of us noted that the middle codon base was itself a good predictor of amino acid hydrophobicities ($\Delta G_{w>c}$). In fact, the full anticodon code (*SI Appendix*, Table S1) selects the appropriate $\Delta G_{w>c}$ value without ambiguity, i.e., $R^2 = 1.00$ (*SI Appendix*, Table S2 and Fig. 2E). Consistency with the 20 amino acid training set gives an RMS error of 0.14 kcal/mol. The RMS relative error of Sec and Pyl predictions is 0.22 (*SI Appendix*, Table S5). The acceptor stem code (*SI Appendix*, Table S3) specifies $\Delta G_{v>c}$ with similar precision (Fig. 2B and *SI Appendix*, Table S4; $R^2 = 1.00$; RMS error = 0.39 kcal/mol). The RMS relative error of Sec and Pyl predictions is 0.13 (*SI Appendix*, Table S5).

Acceptor stem coding for $\Delta G_{w>c}$ (Fig. 2C) and anticodon coding for $\Delta G_{v>c}$ (Fig. 2E) both fail to predict the behavior of Sec and Pyl. The RMS relative errors of Sec and Pyl are 2.74 and 1.20, respectively, which are an order of magnitude larger than

Fig. 2. tRNA coding of amino acid properties. (*A*) Binary representation of tRNA coding showing the acceptor stem (green) and anticodon (red). (*B–E*) Correlations between experimental values for $\Delta G_{w>c}$ (*C* and *E*) and $\Delta G_{v>c}$ (*B* and *D*) and those calculated from the best regression models (II). *B–E* are arranged as a 2^2 factorial design for the two tRNA coding regions (down the vertical) and the two physicochemical properties (across the horizontal). Coefficients trained on the 20 canonical amino acids (*SI Appendix*, Tables S2 and S4) were used to predict values for Sec and Pyl for cross-validation. Lower right-hand corners of *B–E* show RMS relative errors for cross-validation. Plots prepared using JMP (63).

those in Fig. 2 *B* and *E* (0.13 and 0.22). Failure to predict the behavior of the noncanonical amino acids is strong evidence that these two codes are overfitted. We conclude that anticodon and acceptor stem bases specify orthogonal physical properties of the 20 amino acids that direct protein folding. The two complementary coding systems reinforce one another, each enhancing specificity where the other is inadequate.

aaRS Class Contributes Significantly to Both Anticodon and Acceptor-Stem Coding. The acceptor stem coding model for $\Delta G_{v>c}$ requires specifying the groove bound by the synthetase (*SI Appendix*, Table S4). The anticodon coding model for $\Delta G_{w>c}$ requires specifying the aaRS class (*SI Appendix*, Table S2). This distinction arises from the approximate twofold rotation axis relating the acceptor-TΨC and anticodon-D stems (30). Subclass A and B synthetases approaching the acceptor stem from the minor groove also approach the anticodon loop from the minor groove. However, the distinction between the two grooves is unambiguous only in the acceptor stem (31). Subclass IC and IIC synthetases approach one of the two strands of the anticodon loop, leaving the distinction between major and minor grooves—but not class—ambiguous. Eliminating the class/groove distinction substantially compromises both models (*SI Appendix*, Table S5).

The Two tRNA Coding Regions Also Discriminate Between Different Subsets of Seven Categorical Side-Chain Descriptors. The acceptor stem distinguishes between β-branched and nonbranched side-chains but fails to identify aromatic, positively charged, or amide-containing side-chains correctly (Table 1). The anticodon identifies these correctly but fails to distinguish β-branched from nonbranched side-chains. Bases of both the acceptor stem and anticodon identify carboxylate side-chains correctly, suggesting that such side-chains may have played an unusually important role in both early and late stages of protein evolution (see below).

Side-chain–Water Interactions Are Systematically Less Favorable at Higher Temperatures, Without Changing the Anticodon Code. There is widespread, but not universal agreement that life arose soon after the earth cooled sufficiently to support liquid water (32, 33). Temperature changes produce significant effects on amino acid side-chain hydrophobicities (28) (*SI Appendix*, Table S6). All side-chains enter the hydrocarbon phase by a variable amount more at 100 °C than they do at 25 °C. As a consequence, three groups of amino acids (Met, Ala; Cys, Trp; and His, Glu,

Gln, Lys) are ordered slightly differently at 25° and 100 °C (highlighted by colored backgrounds in *SI Appendix*, Table S6).

Remarkably, the same coefficients for the anticodon code (Fig. 2*E*) predict $\Delta G_{w>c}$ at all temperatures (R^2 values are ~1.0; the models predict decreasing $\Delta G_{w>c}$ for Sec and Pyl at higher temperatures). Coefficients for effects of the anticodon bases are nearly temperature independent. Adjusting the intercepts and coefficients for the class distinction accommodates different temperatures without degrading the correlation (*SI Appendix*, Figs. S3 and S4).

Discussion

Here we examine in further detail the relationship noted earlier (10) between amino acid physical chemistry, protein folding, and the genetic code in light of the comprehensive database of tRNA identity elements in the acceptor stem (11) and growing evidence that protein synthesis emerged first using urzyme-like synthetases that recognized only the acceptor stem (17). The simplifying assumption that tRNA identity elements function as a binary—on/off—digital code allowed us to ignore numerous sources of ambiguity, including the effects of base modification and some evolutionary changes in plastid identity elements (34) that were undoubtedly important to the detailed evolution of the code. We find a clear distinction between the

Table 1. Categories coded by tRNA Acceptor stems and anticodons

Functionality	Acceptor stem			Anticodon		
	R^2	df	Coded	R^2	df	Coded
Carboxylate	*1*	*12*	YES	*0.98*	*10*	YES
Aliphatic	*0.99*	*8*	YES	*1*	*8*	YES
β-Branched	*0.99*	*9*	YES	**0.67**	**14**	NO
Aromatic	**0.72**	**13**	NO	*1*	*7*	YES
Charged	**0.58**	**13**	NO	*1*	*3*	YES
Basic	**0.39**	**15**	NO	*1*	*10*	YES
Amide	**0.66**	**16**	NO	*1*	*4*	YES

Complete (italic) vs. poor (bold) coding of categorical amino acid properties. R^2, the squared correlation coefficient, assumes the categorical variables are continuous, and df is the difference between the number of data (i.e., 20 amino acids) and the number of predictors used to estimate coefficients.

amino acid properties sought in aaRS recognition of tRNA acceptor stems and their anticodons.

The aaRS Class Distinction. The two aaRS classes activate sets of 10 amino acids that differ in their median sizes and hydrophobicities (16) (Fig. 1). Side-chain hydrophobicity has long been considered an essential requirement for well-packed globular structures (35). Fig. 1 and *SI Appendix*, Fig. S2 show that side-chain size also contributes systematically, and synergistically, to the ASA in folded proteins. The enhanced separation between the median ASAs of class I and II amino acids (Fig. 1*C*) suggests a potential selective advantage for the amino acid classes. Class I amino acids allowed formation of nonpolar cores and class II amino acids populated the surfaces of globular proteins. The linkage between classes arising from their sense/antisense ancestry (13, 36, 37) would be expected to simplify the search for reduced amino acid alphabets that may have been used during early protein evolution, leading to the universal genetic code.

Independent tRNA Coding Strategies Suggest Distinct Stages of Genetic Coding. Dual coding for amino acids by tRNA acceptor stem and anticodon bases correlates strongly with experimental values for two linearly independent branches of the thermodynamic cycle of vapor to solvent transfers, $K_{v>c}$ and $K_{v>w}$ (*SI Appendix*, Fig. S1). That cycle (19, 28) affords a comprehensive experimental description of how the 20 amino acids direct folding (Fig. 1 *C–E*), supporting the view that acceptor stem and anticodon bases compose full, complementary, and independent, specifications for the 20 canonical amino acids by coding, respectively, for size and polarity.

Danchin (38) raised the following question: Do contemporary genomes constitute archives from which we can hope to extract reasonable histories? Or, are they "palimpsests" (39) in which clues to ancestral structures have been overwritten by entirely new algorithms? He suggests that the sizeable number of genes that persist over all or most organisms supports the former conclusion. Urzyme catalysis extends that argument to molecular levels, in which contemporary aaRS and tRNAs represent archives of ancestral relationships.

The proportion of variance explained by coding elements (*SI Appendix*, Figs. S5 and S6) may reflect the order in which they became part of the code (24) (Fig. 3). The anticodon base predictor selected first in stepwise searches is whether the middle base is a pyrimidine (Y) or a purine (R), accounting for 48% of the variance in $\Delta G_{w>c}$. The second selected predictor is the three-way interaction class × middle base G/A × third base G/A, which implies three new main effects and three new two-way effects. Together, these eight predictors raise the predicted variance to 92%. Turning to the acceptor stem, the two-way interaction between the discriminator base (G/A) and base 1 Y/R, is selected first, accounting for 51% of the variance in $\Delta G_{v>c}$ and confirming the importance of the discriminator base (11, 40). Adding the two-way interaction between bases 70(G/A) and 1(Y/R) accounts for 85% of the variance in size.

The Operational RNA Code in the Acceptor-Stem (17) Can Interpret Genetic Information Without the Anticodon. The acceptor stem code is remarkable in two ways. First, it constitutes an unexpectedly complete code for the 20 canonical amino acids. Second, it encodes the sizes of the amino acids more effectively than their solvent transfer free energies. We interpret Fig. 2 *B* and *E* and Table 1 in terms of the likelihood that protein translation was once driven by urzyme-like synthetases that recognized only ancestral tRNA acceptor stems. The fidelity with which the acceptor stem code might have been interpreted is now accessible to experiment because the modular components—aaRS urzymes and tRNA acceptor stem helices—have now both been characterized (13–16, 41, 42). Preliminary specificity ratios $[(k_{cat}/K_M)_{(cognate)}/(k_{cat}/K_M)_{(noncognate)}]$ for activation of cognate vs. all 19 noncognate amino acids by LeuRS and HisRS urzymes (13, 15, 36, 43) suggest that they prefer amino acids from their own class by ~-1 kcal/mol, or $\sim20\%$ of -5.5 kcal/mol based on the present day specificity ratio of 10^{-4} (44). In due course, k_{cat}/K_M values for aminoacylation of different tRNAs by urzymes will presumably be determined. For the present, these data confirm that proteins synthesized according to the operational code were probably statistical ensembles (45), an idea suggested by biological experiments with mutant aaRS (46–48). Enhanced specificity and catalytic proficiency presumably emerged later as aaRS assimilated anticodon recognition into cognate aaRS–tRNA pairs and gradually assumed their modern forms.
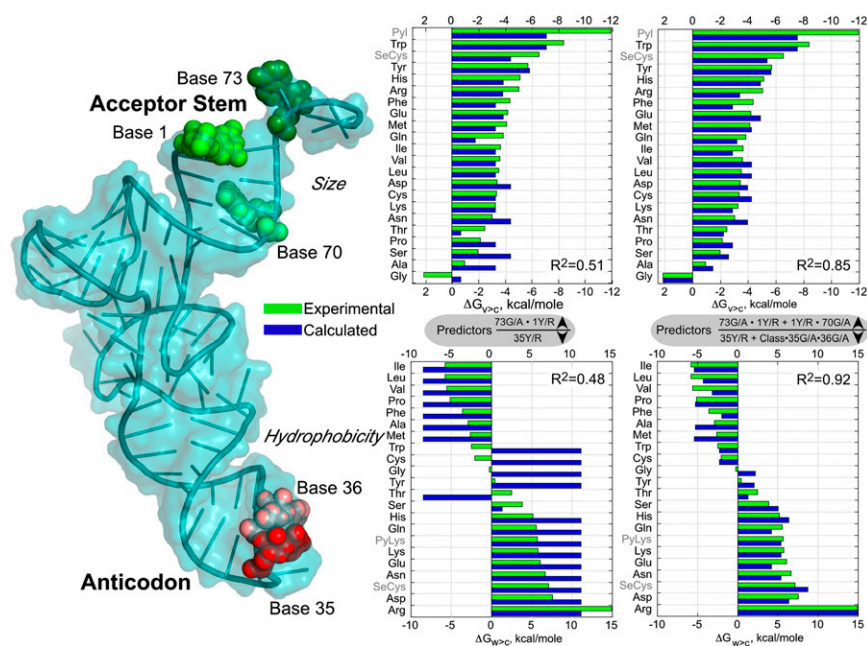
Fig. 3. Orthogonal transfer free energies predicted by the most important bases in tRNA acceptor stems and anticodons. tRNA bases identified by the first two steps in stepwise multiple regression—acceptor stem bases 1, 70, the discriminator base 73, and anticodon bases 35 (middle) and 36 (third), are shown as colored spheres inside the transparent tRNA surface. Histograms show, respectively, experimental (green) and calculated (blue) values of $\Delta G_{v>c}$ for the acceptor stem and $\Delta G_{w>c}$ for the anticodon. Predictors are indicated between the upper and lower histogram for the first two identified predictors.

The Acceptor-Stem Code Is Consistent with Preserving Structural Backbone Complementarity Between Extended β-Polypeptide and RNA Double Helices (49–51). Acceptor stem coding may have conferred a selective advantage by distinguishing smaller from larger amino acids and identifying β-branched, aliphatic, and carboxylate side-chains. Fig. 4 shows how only small side-chains fit between peptide and RNA bases, whereas side-chains pointing away from the RNA are not so constrained. Coding of amino acid size could thus have helped preserve these patterns in transitions from a proposed direct, stereochemical specification (49, 50) to triplet coding. The conformational propensity of β-branched side-chains could similarly have favored β-secondary structures (52).

Lysine and arginine might be expected to have played roles in forming complementary ion pairs with nucleic acid phosphate groups, but the acceptor stem code does not discriminate them from other amino acids. At first, the distinctive coding for carboxylate side-chains seems paradoxical. However, carboxylate groups are the dominant ligand for Mg^{2+} ions (53) and may have coordinated other divalent metals, such as ferrous iron (54). Coordination of Mg^{2+} ions has been cited as potentially useful for countering the metal-catalyzed hydrolysis of RNA (55) and hence for the emergence of stable oligonucleotides (56). Moreover, Mg^{2+} ions are now the dominant divalent metals in transferases and ligases (57), and these catalysts are, by a considerable margin, the most important activities for the origin of replicating systems. Finally, peptide insolubility is a substantial problem that might have been even more important in a peptide/RNA world. Recent studies (58) associated carboxylate side-chains uniquely with increased solubility, and early codes might also have selected carboxylate groups for that reason.

Although we are not in a position to address the question of how aminoacylated tRNA acceptor stems might have been aligned in accordance with a primordial mRNA without anticodons, Rodin and Ohno suggested that reconstructed tRNA acceptor stems display evidence of complementary sequences (59). Our results revive the possibility that such complementarity and/or lateral-loop-loop base-pairing (60) might have aligned acylated acceptor stems, anticipating the assembly of peptides according to a message.

Conclusions

We suggest that genetic coding evolved in distinct stages. Initial discrimination on the basis of size may have allowed coding by tRNA acceptor stems to ensure that the earliest peptides were β-structures with alternating large and small side-chains, to interact with RNA, and only later encoded globular conformations with greater catalytic activity. The earliest peptides may have included the unstructured peptide tails that stabilize ribosomal RNAs (61). Systematic analysis of their amino acid compositions and multiple sequence alignments may reveal patterns related to properties suggested here by tRNA acceptor stem coding. Evolution of tRNA identity elements is also important for understanding idiosyncratic coding, for example, as in plastids (34). The order in which predictors emerge in the stepwise regressions discussed above is similar, but not identical to, the series of decisions by which Delarue suggested that genetic coding actually became fixed (24). Although tRNA identity elements have probably been confounded by horizontal gene transfer (62), ancestral tRNA sequence reconstruction may clarify further how identity elements and the synthetase Class recognition evolved.

Methods

Vapor Phase > Solvent Transfer Free Energies and Their Temperature Dependences. $\Delta G_{w>c}$ values (19) were redetermined for this study (28) and used to recalculate the corresponding values for $\Delta G_{v>c}$ as described in *SI Appendix*, section 1 and *SI Appendix*, Table S6.
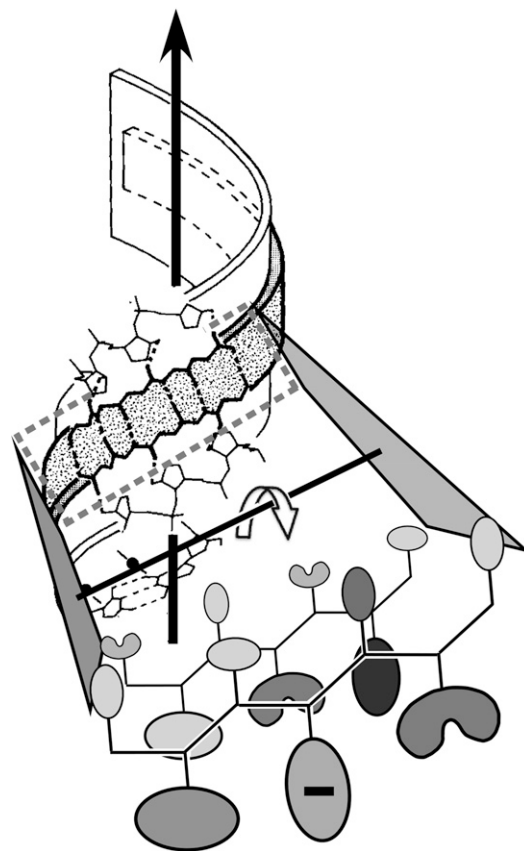
Fig. 4. Possible relevance of size, β-branching, and carboxylates to the operational RNA code. For ancient β-hairpins to interact with double-stranded RNA as envisioned by Carter and Kraut (50), large side-chains would necessarily face away from the RNA minor groove. β-branched side-chains (reentrant curves; Thr, Val on the inward face; Ile on the outer face) enhance β-structure formation. Carboxylate side-chains facing outward could coordinate divalent cations for catalysis, to limit RNA degradation, or to enhance solubility.

Binary Coding by tRNA Bases. tRNA bases are each assigned two bits, anticodon bases being taken in the order base 2, then base 3, and then base 1 to reflect their relative importance. The first bit denotes whether the corresponding codon–anticodon interaction forms three hydrogen bonds (i.e., G or C = 1) or two (A or U = −1); the second denotes whether the base is a pyrimidine (i.e., Y = 1) or a purine (R = −1). For example, isoleucine anticodons RAU generate a seven-term vector {1 −1 −1 −1 1 0 −1}, that begins with the class (1), followed by two two-bit elements each for base 2, then base 3, and finally base 1. Here, 0 represents ambiguity associated with the first base, which can be either G or A. Acceptor stem coding proceeds in analogous fashion for the eight bases in the stem plus the discriminator base D (usually 73) (40) and the groove, for a total of 19 bits of potential information (*SI Appendix*, Table S3). Note that synthetases read coding bases in either of two ways, because class I and II aaRSs bind to opposite sides of the acceptor stem. In general, class II enzymes bind to the major groove and class I to the minor groove. However, for subclasses Ic and IIc, this situation is reversed (31).

Multivariate Regression Analysis. Multiple regression was performed using JMP (63). We estimated $\Delta G_{v>c}$ for the two additional amino acids from their known masses, using the correlation between mass and $\Delta G_{v>c}$ for the 20 canonical amino acids (details in *SI Appendix*, sections 3 and 4 and Figs. S5 and S6).

1. Eriani G, Delarue M, Poch O, Gangloff J, Moras D (1990) Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature* 347(6289):203–206.
2. Cusack S, Berthet-Colominas C, Härtlein M, Nassar N, Leberman R (1990) A second class of synthetase structure revealed by X-ray analysis of *Escherichia coli* seryl-tRNA synthetase at 2.5 A. *Nature* 347(6290):249–255.
3. Min B, et al. (2003) Protein synthesis in Escherichia coli with mischarged tRNA. *J Bacteriol* 185(12):3524–3526.
4. Deniziak M, et al. (2007) Deinococcus glutaminyl-tRNA synthetase is a chimer between proteins from an ancient and the modern pathways of aminoacyl-tRNA formation. *Nucleic Acids Res* 35(5):1421–1431.
5. Giegé R, Eriani G (2014) *Transfer RNA Recognition and Aminoacylation by Synthetases in eLS* (John Wiley & Sons, Ltd, Chichester, UK), pp 1–18.
6. Crick FHC (1968) The origin of the genetic code. *J Mol Biol* 38(3):367–379.
7. Woese C (1969) Models for the evolution of codon assignments. *J Mol Biol* 43(1): 235–240.
8. Freeland SJ, Hurst LD (1998) The genetic code is one in a million. *J Mol Evol* 47(3): 238–248.
9. Koonin EV, Novozhilov AS (2009) Origin and evolution of the genetic code: The universal enigma. *IUBMB Life* 61(2):99–111.
10. Wolfenden RV, Cullis PM, Southgate CCF (1979) Water, protein folding, and the genetic code. *Science* 206(4418):575–577.
11. Giegé R, Sissler M, Florentz C (1998) Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res* 26(22):5017–5035.
12. Sanderson LE, Uhlenbeck OC (2007) The 51-63 base pair of tRNA confers specificity for binding by EF-Tu. *RNA* 13(6):835–840.
13. Li L, Francklyn C, Carter CW, Jr (2013) Aminoacylating urzymes challenge the RNA world hypothesis. *J Biol Chem* 288(37):26856–26863.
14. Li L, Weinreb V, Francklyn C, Carter CW, Jr (2011) Histidyl-tRNA synthetase urzymes: Class I and II aminoacyl tRNA synthetase urzymes have comparable catalytic activities for cognate amino acid activation. *J Biol Chem* 286(12):10387–10395.
15. Pham Y, et al. (2010) Tryptophanyl-tRNA synthetase Urzyme: A model to recapitulate molecular evolution and investigate intramolecular complementation. *J Biol Chem* 285(49):38590–38601.
16. Pham Y, et al. (2007) A minimal TrpRS catalytic domain supports sense/antisense ancestry of class I and II aminoacyl-tRNA synthetases. *Mol Cell* 25(6):851–862.
17. Schimmel P, Giegé R, Moras D, Yokoyama S (1993) An operational RNA code for amino acids and possible relationship to genetic code. *Proc Natl Acad Sci USA* 90(19): 8763–8768.
18. Giegé R (1972) Recherches sur la spécificité de reconnaissance des acides ribonucléiques de transfert par les aminoacyl-tRNA synthétases [Study on the specificity of recognition of transfer ribonucleic acids by aminoacyl-tRNA synthetases]. PhD Thesis. *Biological Chemistry* (Université Louis Pasteur, Strasbourg, France).
19. Radzicka A, Wolfenden R (1988) Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* 27(5):1664–1670.
20. Hine J, Mookerjee P (1975) Structural effects on rates and equlibriums. XIX. Intrinsic hydrophilic character of organic compounds. Correlations in terms of structural contributions. *J Org Chem* 40(3):292–298.
21. Creighton TE (1984) *Proteins: Structures and Molecular Properties* (Freeman, New York).
22. Lee B, Richards FM (1971) The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol* 55(3):379–400.
23. Chothia C (1976) The nature of the accessible and buried surfaces in proteins. *J Mol Biol* 105(1):1–12.
24. Delarue M (2007) An asymmetric underlying rule in the assignment of codons: Possible clue to a quick early evolution of the genetic code via successive binary choices. *RNA* 13(2):161–169.
25. Rodin SN, Rodin AS (2008) On the origin of the genetic code: Signatures of its primordial complementarity in tRNAs and aminoacyl-tRNA synthetases. *Heredity (Edinb)* 100(4):341–355.
26. Belrhali H (1996) Détermination par Cristallographie aux Rayons X des Mécanismes de Formation du Seryl-Adenylate et du bis(5′-Adenosyl) Tetraphosphate par la Seryl-aRNT Synthetase de Thermus Thermophilus. [Determination by X-ray Crystallographie of the Mechanisms by which Seryl-Adenylate and bis(5′-Adenosyl) Tetraphosphate are formed by the Seryl-tRNA Synthetase from Thermus Thermophilus]. PhD thesis. *Sciences Biologiques Fondamentales et Appliquées* (European Synchrotron Radiation Facility, Grenoble, France).
27. Moelbert S, Emberly E, Tang C (2004) Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins. *Protein Sci* 13(3):752–762.
28. Wolfenden R, Lewis CA, Jr, Yuan Y, Carter CW, Jr (2015) Temperature dependence of amino acid hydrophobicities. *Proc Natl Acad Sci USA* 112:7484–7488.
29. Neter J, Wasserman W (1974) *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs* (Richard D. Irwin, Homewood, IL).
30. Kim S (1975) Symmetry recognition hypothesis model for tRNA binding to aminoacyl-tRNA synthetase. *Nature* 256(5519):679–681.
31. Ribas de Pouplana L, Schimmel P (2001) Two classes of tRNA synthetases suggested by sterically compatible dockings on tRNA acceptor stem. *Cell* 104(2):191–193.
32. Miller SL, Lazcano A (1995) The origin of life—Did it occur at high temperatures? *J Mol Evol* 41:689–692.
33. Gaucher EA, Govindarajan S, Ganesh OK (2008) Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* 451(7179):704–707.
34. Igloi GL, Leisinger A-K (2014) Identity elements for the aminoacylation of metazoan mitochondrial tRNA(Arg) have been widely conserved throughout evolution and ensure the fidelity of the AGR codon reassignment. *RNA Biol* 11(10):1313–1323.
35. Kauzmann W (1959) Some factors in the interpretation of protein denaturation. *Adv Protein Chem* 14:1–63.
36. Carter CW, Jr, et al. (2014) The Rodin-Ohno hypothesis that two enzyme superfamilies descended from one ancestral gene: An unlikely scenario for the origins of translation that will not be dismissed. *Biol Direct* 9:11.
37. Chandrasekaran SN, Yardimci GG, Erdogan O, Roach J, Carter CW, Jr (2013) Statistical evaluation of the Rodin-Ohno hypothesis: Sense/antisense coding of ancestral class I and II aminoacyl-tRNA synthetases. *Mol Biol Evol* 30(7):1588–1604.
38. Danchin A (2007) Archives or palimpsests? Bacterial genomes unveil a scenario for the origin of life. *Biol Theory* 2(1):1–10.
39. Benner SA, Ellington AD, Tauer A (1989) Modern metabolism as a palimpsest of the RNA world. *Proc Natl Acad Sci USA* 86(18):7054–7058.
40. Crothers DM, Seno T, Söll G (1972) Is there a discriminator site in transfer RNA? *Proc Natl Acad Sci USA* 69(10):3063–3067.
41. Francklyn C, Musier-Forsyth K, Schimmel P (1992) Small RNA helices as substrates for aminoacylation and their relationship to charging of transfer RNAs. *Eur J Biochem* 206(2):315–321.
42. Francklyn C, Schimmel P (1989) Aminoacylation of RNA minihelices with alanine. *Nature* 337(6206):478–481.
43. Carter CW, Jr (2014) Urzymology: experimental access to a key transition in the appearance of enzymes. *J Biol Chem* 289(44):30213–30220.
44. Fersht AR (1999) *Structure and Mechanism in Protein Science* (W. H. Freeman and Company, New York).
45. Woese CR (1965) On the evolution of the genetic code. *Proc Natl Acad Sci USA* 54(6): 1546–1552.
46. Giegé R, et al. (2012) Structure of transfer RNAs: Similarity and variability. *Wiley Interdiscip Rev RNA* 3(1):37–61.
47. White SH (1994) The evolution of proteins from random amino acid sequences: II. Evidence from the statistical distributions of the lengths of modern protein sequences. *J Mol Evol* 38(4):383–394.
48. Pezo V, et al. (2004) Artificially ambiguous genetic code confers growth yield advantage. *Proc Natl Acad Sci USA* 101(23):8593–8597.
49. Carter CW (1975) Cradles for molecular evolution. *New Scientist* March 27:784–787.
50. Carter CW, Jr, Kraut J (1974) A proposed model for interaction of polypeptides with RNA. *Proc Nat Acad Sci USA* 71(2):283–287.
51. Carter CW, Jr (2015) What RNA World? Why a Peptide/RNA Partnership Merits Renewed Experimental Attention. *Life (Basel)* 5(1):294–320.
52. Muñoz V, Serrano L (1994) Intrinsic secondary structure propensities of the amino acids, using statistical φ-ψ matrices: Comparison with experimental scales. *Proteins Struct Funct Bioinform* 20(3):301–311.
53. Glusker JP, Katz AK, Bock CW (1999) Metal ions in biological systems. *Rigaku J* 16(2): 8–16.
54. Athavale SS, et al. (2012) RNA folding and catalysis mediated by iron (II). *PLoS ONE* 7(5):e38024.
55. AbouHaidar MG, Ivanov IG (1999) Non-enzymatic RNA hydrolysis promoted by the combined catalytic activity of buffers and magnesium ions. *Z Naturforsch C* 54(7-8): 542–548.
56. Szostak J (2012) The eightfold path to non-enzymatic RNA replication. *J Sys. Chem.* 3:2.
57. Andreini C, Bertini I, Cavallaro G, Holliday GL, Thornton JM (2008) Metal ions in biological catalysis: From enzyme databases to general principles. *J Biol Inorg Chem* 13(8):1205–1218.
58. Kramer RM, Shende VR, Motl N, Pace CN, Scholtz JM (2012) Toward a molecular understanding of protein solubility: Increased negative surface charge correlates with increased solubility. *Biophys J* 102(8):1907–1915.
59. Rodin S, Rodin A, Ohno S (1996) The presence of codon-anticodon pairs in the acceptor stem of tRNAs. *Proc Natl Acad Sci USA* 93(10):4537–4542.
60. Henderson BS, Schimmel P (1997) RNA-RNA interactions between oligonucleotide substrates for aminoacylation. *Bioorg Med Chem* 5(6):1071–1079.
61. Fox GE (2010) Origin and evolution of the ribosome. *Cold Spring Harb Perspect Biol* 2:a003483.
62. Ardell DH, Andersson SGE (2006) TFAM detects co-evolution of tRNA identity rules with lateral transfer of histidyl-tRNA synthetase. *Nucleic Acids Res* 34(3):893–904.
63. SAS (2007) *JMP Statistics and Graphics Guide* (SAS Institute, Cary, NC).