RNA
A PUBLICATION OF THE RNA SOCIETY

Revisiting the operational RNA code for amino acids: Ensemble attributes and their implications

Shaul Shaul, Dror Berel, Yoav Benjamini, et al.

RNA 2010 16: 141-153 originally published online December 1, 2009
Access the most recent version at doi:10.1261/rna.1745910

Revisiting the operational RNA code for amino acids: Ensemble attributes and their implications

SHAUL SHAUL,[1] DROR BEREL,[2] YOAV BENJAMINI,[3] and DAN GRAUR[1,4]

[1]Department of Zoology, Tel Aviv University, Tel Aviv 69978, Israel
[2]Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, USA
[3]Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel
[4]Department of Biology and Biochemistry, University of Houston, Houston, Texas 77004, USA

ABSTRACT

It has been suggested that tRNA acceptor stems specify an operational RNA code for amino acids. In the last 20 years several attributes of the putative code have been elucidated for a small number of model organisms. To gain insight about the ensemble attributes of the code, we analyzed 4925 tRNA sequences from 102 bacterial and 21 archaeal species. Here, we used a classification and regression tree (CART) methodology, and we found that the degrees of degeneracy or specificity of the RNA codes in both Archaea and Bacteria differ from those of the genetic code. We found instances of taxon-specific alternative codes, i.e., identical acceptor stem determinants encrypting different amino acids in different species, as well as instances of ambiguity, i.e., identical acceptor stem determinants encrypting two or more amino acids in the same species. When partitioning the data by class of synthetase, the degree of code ambiguity was significantly reduced. In cryptographic terms, a plausible interpretation of this result is that the class distinction in synthetases is an essential part of the decryption rules for resolving the subset of RNA code ambiguities enciphered by identical acceptor stem determinants of tRNAs acylated by enzymes belonging to the two classes. In evolutionary terms, our findings lend support to the notion that in the pre-DNA world, interactions between tRNA acceptor stems and synthetases formed the basis for the distinction between the two classes; hence, ambiguities in the ancient RNA code were pivotal for the fixation of these enzymes in the genomes of ancestral prokaryotes.

Keywords: cryptography; operational RNA code; amino acids

INTRODUCTION

In aminoacylation, the first step of protein synthesis, each aminoacyl-tRNA synthetase (aaRS) must recognize its cognate tRNA, discriminating against all the others, ligate covalently the corresponding amino acid to the 3′ terminus of the acceptor stem, and edit out misacylated amino acids. The accuracy of protein synthesis rests not only on the high intrinsic fidelity of the aminoacylation process, but also on competition between different synthetases for a given tRNA (Sherman et al. 1992).

Synthetases (aaRSs) are among the most ancient enzymes, ostensibly as ancient as the genetic code (Ribas de Pouplana and Schimmel 2001b). Despite performing exactly the same function (tRNA aminoacylation), they belong to two structurally unrelated protein families (Eriani et al. 1990). Class-1 enzymes differ from class-2 enzymes in secondary structure, sequence motifs at the active site, the side of the tRNA acceptor stem onto which they dock, and the domains responsible for editing out misacylated amino acids with steric and/or biochemical similarity. Members of class-1 approach the tRNA acceptor stem from the minor groove side, dock on the 5′ branch, and acylate at the 2′ hydroxyl group of the A76 terminus; whereas those of class-2 approach from the major groove side, dock on the 3′ branch, and acylate at the 3′-OH of the terminal adenine (Giegé et al. 1974; Cavarelli and Moras 1993; Arnez et al. 1995; Fersht 1998; Nureki et al. 1998; Fukai et al. 2000; Beebe et al. 2003, 2004). Only PheRS, although a class-2, attaches the amino acid to the 2′-OH (Goldgur et al. 1997). For 19 of the 20 primary amino acids, there exists only one specific enzyme ligating its cognate amino acid to tRNAs bearing anticodons corresponding to that amino acid. For lysine, there are two synthetases, LysRS1 and LysRS2, featuring the respective characteristics of class-1 and class-2

Reprint requests to: Dan Graur, Department of Biology and Biochemistry, University of Houston, Houston, TX 77004, USA; e-mail: dgraur@uh.edu; fax: (713) 743-2636.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.1745910.

synthetases (Ibba et al. 1997, 1999). Sequence analysis of more than 700 synthetases indicated that the two distinct families have no common ancestor (Schimmel and Ribas de Pouplana 2001). While no definitive explanation for the origin of this class distinction exists, several conjectures have been advanced (Schimmel and Ribas de Pouplana 2001; Ribas de Pouplana and Schimmel 2001b,c; Rodin and Rodin 2008).

The two-dimensional tRNA cloverleaf molecule (Marck and Grosjean 2002) consists of two halves, the top part of which, the minihelix, consists of the acceptor stem and the TψC arm. It is thought that the top and bottom parts of tRNA have originated by duplication (Möller and Janssen 1990; Di Giulio 1995; Rodin and Rodin 2008). The most parsimonious way for maintaining the fidelity of the genetic code should have been through the direct recognition of tRNA anticodons by cognate synthetases. Yet, the anticodon is not always the principal determinant for aminoacylation (Shimada et al. 2001; Weygand-Durasevic et al. 2002; Jones et al. 2008), and in some cases, no physical contact is made between the aaRS and the anticodon (Park and Schimmel 1988). A set of "identity elements" classified as "determinants" and "antideterminants" (Giegé et al. 1993), distributed over the stems and loops of the tRNA molecule, participate in its interaction with enzymes and the ribosome during maturation, modification, biological function, and degradation (Wolfson et al. 2001). Determinants facilitate aminoacylation; antideterminants impede it. Their combined effect ensures that a particular tRNA is a substrate for its cognate synthetase and not a substrate for all the other synthetases present in a cell (McClain 1993; Nureki et al. 1994; Giegé et al. 1998; McClain et al. 1998; Beuning and Musier-Forsyth 1999; Fender et al. 2004). An updated list of determinants based on a compilation by Giegé et al. (1998) is shown in Supplemental Table S1.

In a series of groundbreaking experiments, Paul Schimmel and colleagues identified nucleotides and specific structural features within and outside the tRNA acceptor stem that affect the efficiency of aminoacylation (Francklyn and Schimmel 1989, 1990; Martinis and Schimmel 1993; Schimmel et al. 1993; Hamann and Hou 1995; Schimmel 1995; Saks and Sampson 1996; Di Giulio 1997; Musier-Forsyth and Schimmel 1999; Ribas de Pouplana and Schimmel 2001a; Shimada et al. 2001; Weygand-Durasevic et al. 2002). In particular, determinants at various positions along the acceptor stem that were found to be directly responsible for proper tRNA aminoacylation led to the concept of an "operational RNA code for amino acids" (Schimmel et al. 1993; Schimmel 1995), indicating that the relationship between amino acids and tRNAs is encrypted not once but twice in the molecule. In contrast to the universal nature of the genetic code, the RNA code is often species-specific: a synthetase from one source will not acylate its cognate tRNA from another (Nair et al. 1997; Lovato et al. 2001; Xu et al. 2001) and varies in evolution due to coadaptations of the contact residues between synthetases and acceptor stems of cognate tRNAs (Stehlin et al. 1998). While experiments with model organisms indicated that the N73 "discriminator" base (Grothers et al. 1972) and the first four base pairs of the tRNA acceptor stem and of synthetic RNA minihelices may be sufficient for specific aminoacylations in a number of model organisms (Francklyn and Schimmel 1989; Martinis and Schimmel 1993, 1995; Hamann and Hou 1995; Schimmel and Ribas de Pouplana 1995; Saks and Sampson 1996; Musier-Forsyth and Schimmel 1999), the results of other experiments demonstrated that there are cases for which (1) aaRS-tRNA recognition cannot be reduced to isolated structural elements, but, rather, the tRNA acceptor stem is being recognized as a unit (Choi et al. 2002); and (2) recognition includes determinants other than the acceptor stem (McClain et al. 1998). To account for major aminoacylation determinants outside the acceptor stem, the term "generalized operational RNA code for amino acids" was coined (Schimmel et al. 1993). It encompasses all nucleotides in the tRNA, exclusive of the anticodon, that affect the efficiency of specific aminoacylation. The two most prominent examples are the A20 nucleotide in the D-loop of $tRNA^{Arg}$ and the nucleotides in the long variable arm of $tRNA^{Ser}$, which are recognized by unique domains in the N termini of ArgRS and SerRS, respectively (Cavarelli et al. 1998; Giegé et al. 1998; Hendrickson 2001; Shimada et al. 2001; Weygand-Durasevic et al. 2002).

While some quantitative and qualitative features of the operational RNA code, for example, its species specificity and its variability in evolution, have been experimentally deciphered for a rather small sample of model organisms, a comprehensive understanding of the cryptographic rules governing the relationship between the RNA code and synthetases and the ensemble characteristics of the operational RNA code is still lacking. We believe that these will emerge by investigating the code attributes with the aid of data-mining statistical tools on a taxonomic scale that is much larger than practical for experimental tackling. The availability of hundreds of completely sequenced prokaryotic genomes allows us to address these issues and speculate specifically whether attributes of the RNA code may have been the putative agents for the enigmatic origin of the two distinct classes of synthetases in the genomes of ancestral prokaryotes, in line with Schimmel and Ribas de Pouplana (2001), who surmised that interactions of the catalytic domains of synthetases with tRNA acceptor stems may have formed the basis for the distinct classes of these enzymes. We emphasize the exploratory nature of our study; it generates a list of attributes of the operational RNA code for amino acids and some of their cryptographic relationships with aminoacyl-tRNA synthetases that can serve as hypotheses to be validated or refuted experimentally.

RESULTS

The ensemble attributes of the archaeal operational RNA code as inferred from 879 tRNA acceptor stems are presented in Figure 1. It features 79 paths (sets of rules). For ease of use, the much bigger bacterial tree encompassing 437 rules is provided in tabular format, which also features the copy number of amino acids occupying the terminal nodes (Supplemental Table S4). For the database at hand, these sets of rules define the sequence space of the operational RNA codes for amino acids.

Degeneracy of the code

Two or more paths on the tree may end in the same amino acid. This is indicative of a degeneracy in the operational RNA code, analogous to the degeneracy in the genetic code. For example, in Figure 1, there are two paths leading to glutamic acid (E). The eighth is "IF 73A and 71G and 3Y (C or U) and 72C and 70A, THEN E." The eighteenth is "IF 73A and 71G and 3Y (C or U) and 72C and 70G and 5Y (C or U) and 6G and 68G, THEN E." Thus, the archaeal RNA code for glutamic acid in our sample is twofold degenerate. The degrees of degeneracy in the RNA code are defined as the number of distinct sequences of acceptor stem determinants in the archaeal or bacterial tree corresponding to particular amino acids. They are shown in Table 1.

Multiplicities

Twenty-one out of the 79 terminal nodes in the archaeal tree (Fig. 1) and 62 out of the 437 in the bacterial tree (Supplemental Table S4) harbor multiple amino acids. The occurrence of terminal nodes harboring "multiplicities" indicates two qualitatively distinct features of the operational RNA code. The first points to the existence of taxon-specific alternative operational RNA codes, that is, identical tRNA acceptor stem determinants encrypting different

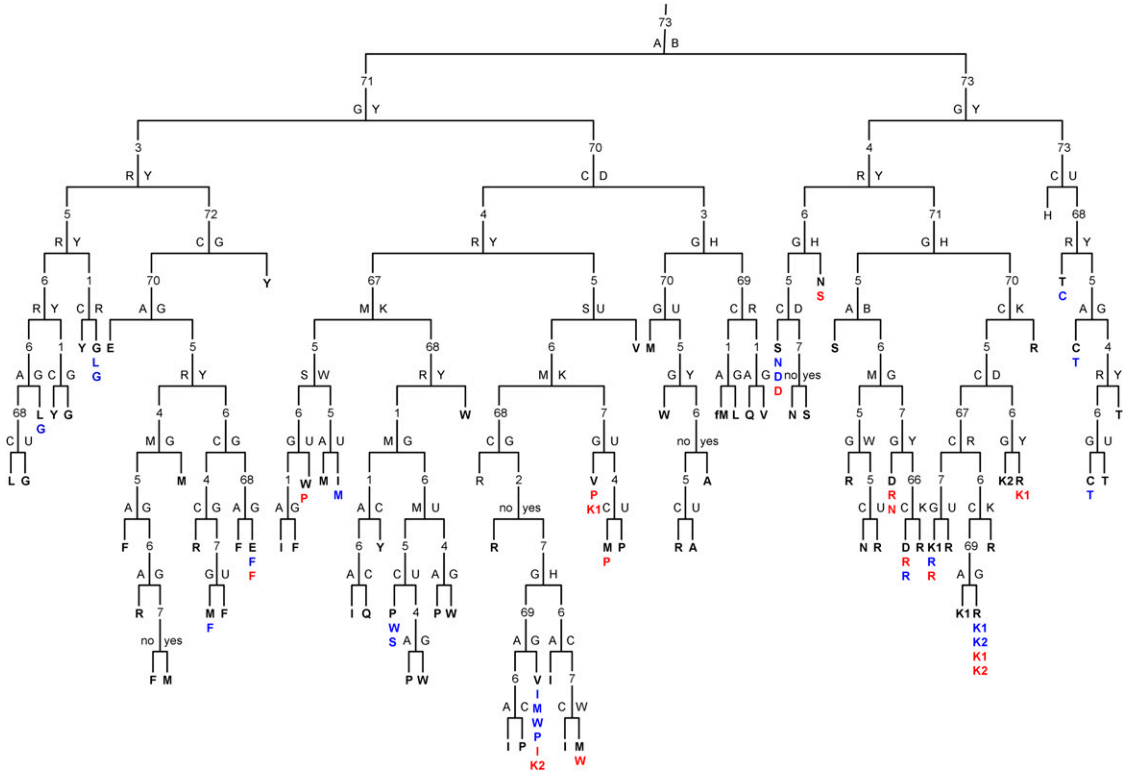

FIGURE 1. Classification and regression tree of domain Archaea. CART generated 79 paths for the 879 tRNA acceptor stems. Unmodified nucleotides at the root of the tree and its nodes are indicated by standard one-letter abbreviations: (A) adenine; (C) cytosine; (U) uracil; (G) guanine; (W, weak bonds) A or U; (S, strong bonds) C or G; (R, purines) A or G; (Y, pyrimidines) C or U; (K, keto) U or G; (M, amino) A or C; (B) C or G or U; (D) A or G or U; (H) A or C or U; (V) A or C or G (Graur and Li 2000). Numbers on vertical branches indicate positions on the acceptor stem. Decision rules are listed at the node below the position number. ("yes/no") Indicates bifurcations determined by the existence of Watson–Crick or non-Watson–Crick pairs, respectively. The tree is rooted at the discriminator nucleotide N73. The amino acids occupying the terminal nodes are denoted by standard one-letter abbreviations (Graur and Li 2000). K1 and K2 indicate lysines acylated to cognate tRNAs by LysRS1 and LysRS2, respectively. The terminal leaves on the tree are referred to in the text and the tables by their sequential number from *left* to *right*; the first leaf harbors amino acid leucine (L); the 79th harbors threonine (T). Twenty-one leaves harbor multiplicities. In these, the predominant amino acid is in black, the minority ones are in red if associated with taxon-specific alternative RNA codes, and blue if associated with ambiguity in the code. The totality of paths on a tree and the content of its terminal nodes constitute the operational RNA code for the data set of all tRNAs in an ensemble of species under study.

TABLE 1. Degrees of degeneracy in the operational RNA codes of Archaea and Bacteria

Amino acid	Number of tRNAs	Degree of degeneracy	Rank (averaged for ties)
Archaea			
His	20	1	1.5
fMet	19	1	1.5
Ala	57	2	4
Gln	37	2	4
Glu	33	2	4
Cys	21	3	6.5
Lys2	8	3	6.5
Asp	20	4	10
Gly	60	4	10
Leu	93	4	10
Tyr	21	4	10
Val	59	4	10
Asn	21	5	14
Ser	78	5	14
Thr	56	5	14
Ile	23	6	16.5
Lys1	28	6	16.5
Phe	21	8	18.5
Trp	18	8	18.5
Pro	57	9	20
Met	37	10	21
Arg	92	18	22
Bacteria			
Ala	219	2	1.5
Cys	106	2	1.5
Gly	254	5	3.5
His	100	5	3.5
fMet	117	6	5
Pro	215	10	6
Lys1	19	15	7.5
Tyr	104	15	7.5
Asp	110	16	9.5
Ser	355	16	9.5
Asn	120	17	11
Gln	146	18	12
Ile	103	20	13
Trp	105	23	14
Phe	108	27	15
Glu	135	29	16
Thr	270	30	17
Met	194	34	18
Lys2	145	36	19
Val	231	39	20
Leu	471	65	21
Arg	384	86	22

amino acids in different organisms. For example, in Figure 1, the 42nd path *7*3A → 71Y → 70C → 4Y → 5S → 6K → 7G leads to a terminal node harboring three types of amino acids: valine (V), proline (P), and lysine (K1). Altogether six amino acids occupy the leaf: four V, one P, and one K1, encoded in determinants in six tRNAs acceptor stems of 4 archaeons (Table 2). Three valines are encoded by the determinants 73A, 4C, 5C, 6G, 7G, 71C, 70C of *Methanothermobacter thermautotrophicus* and *Methanococcus jannaschii*, while the same determinants specify in *Halobacterium sp*. the taxon-specific RNA code variant for proline; one valine is encoded by the determinants 73A, 4U, 5C, 6G, 7G, 71C, and 70C of *Halobacterium sp*., while the same determinants specify in *Thermoplasma acidophilum* the taxon-specific RNA code variant for lysine. Nineteen out of the 21 Archaea in our data set encrypt 37 taxon-specific alternative RNA codes. Among the 102 Bacteria, 63 species encrypt 89 taxon-specific alternative operational RNA codes (Supplemental Table S4; Materials and Methods). Taxon-specific alternative operational RNA codes do not affect the aggregate of unique acceptor stem determinant sequences related to particular amino acids in the data set. They are analogous to the alternative genetic codes (Schultz and Yarus 1996; Jukes and Osawa 1997; Yarus and Schultz 1997; Knight et al. 2001; Santos et al. 2004; Elzanowski and Ostell 2007). The second type of multiplicity arises when identical tRNA acceptor stem determinants encrypt two or more different amino acids in the same organisms. This type of multiplicity points to the existence of an ambiguity in the operation RNA code. For example, in Figure 1, the path 73A → 71G → 3Y → 72C → 70G → 5Y → 6C → 4G → 7G ends in the fifteenth terminal node harboring both methionine (M) and phenylalanine (F). Five species (Supplemental Table S2; Materials and Methods) encode the acceptor stem determinants 73A, 3C, 4G, 5C, 6C, 7G, 72C, 71G, and 70G, encrypting these two amino acid: two species, *Nanoarchaeum equitans* and *Pyrobaculum aerophilum*, encode tRNA acceptor stems specifying only methionine in the RNA code; in contrast, each of the other three archaeons—*Sulfolobus solfataricus*, *Sulfolobus tokodaii*, and *Aeropyrum pernix*—encrypts both methionine and phenylalanine with the same acceptor stem determinants, thereby manifesting ambiguity in the RNA code. We found 15 Archaea encoding 137 tRNA acceptor stems out of 879 (15.6%) with determinants encrypting ambiguous RNA codes for 14 amino acids (Supplemental Table S5). Among the Bacteria, ambiguity in the RNA code is much less prevalent: out of 102 species, only 14 were found encoding 30 tRNA acceptor stems out of 4011 (0.75%) encrypting ambiguous RNA codes for 12 amino acids (Table 3). Ambiguities diminish the aggregate of unique acceptor stems related to particular amino acids in the data set. At present, only one analogy is known in the genetic code; several asporogenic *Candida* species encode serine and leucine by a single polysemous codon (Suzuki et al. 1997; Santos et al. 1999).

Validity of the results

The appropriateness of the classification and regression tree (CART) methodology to capture the structure of a real code was investigated by permutation tests. Each simulation consisted of 1000 runs. The number of inferred rules (paths) and the misclassification rates, calculated by

TABLE 2. Taxon-specific archaeal RNA code variants for proline and lysine

Leaf number[a]	Species	Phylum	Amino acid	tRNA acceptor stem				Anticodon	aaRS class
				N73	N1–N7	N72–N66	(y) WC; (n) NWC		
42	*Methanothermobacter thermautotrophicus*[b]	Euryarchaeota	Val	A	oooCCGG	oCCoooo	yyyyyyy	CAC	1
	Methanothermobacter thermautotrophicus[b]	Euryarchaeota	Val	A	oooCCGG	oCCoooo	yyyyyyy	GAC	1
	Methanococcus jannaschii[b]	Euryarchaeota	Val	A	oooCCGG	oCCoooo	yyyyyyy	UAC	1
	Halobacterium sp[c]	Euryarchaeota	Pro	A	oooCCGG	oCCoooo	yyyyyyy	CGG	2
	Halobacterium sp[b]	Euryarchaeota	Val	A	oooUCGG	oCCoooo	yyyyyyy	CAC	1
	Thermoplasma acidophilum[c]	Euryarchaeota	Lys1	A	oooUCGG	oCCoooo	yyyyyyy	CUU	1

[a]See Figure 1.
[b]Archaeons encoding the acceptor stem sequences with determinants specifying the most common amino acid.
[c]Species encoding acceptor stem determinants specifying taxon-specific alternative RNA code variants. Amino acids are listed by standard three-letter abbreviations. Nucleotides are listed by standard one-letter abbreviations: (A) adenine; (C) cytosine; (U) uracil; (G) guanine. (N73–N66, N1–N7) Nucleotide locations on the tRNA acceptor stem. Zeros indicate acceptor stem locations occupied by nucleotides that are not part of the archaeal operational RNA code. (y) Watson–Crick pairs; (n) non-Watson–Crick pairs.

dividing the copy number of misclassified amino acids by the number of acceptor stems, for the real data and the permutated were compared. The actual archaeal tree for 879 acceptor stems (Fig. 1) consisted of 79 inferred paths. The average number of paths for the trees generated from the permutated data was 208 ± 5. In the archaeal tree there were 21 leaves harboring multiplicities for a total of 94 misclassified amino acids. The misclassification rate in the archaeal tree was, therefore, 94/879 = 10.7%. In comparison, the mean misclassification rate in the trees generated by the permutated data was 62.8% ± 0.6%. Similarly, in the bacterial tree there are 437 paths in comparison to an average of 1238 ± 13 paths in the trees generated by the permutated data. The misclassification rate in Bacteria was 2.8% in comparison to a mean misclassification rate of 55.4% ± 0.3% on the trees generated by the permutated data. For both domains, the differences between the number of paths and misclassification rates is dozens of standard deviations away, that is, statistically significant at incalculable probability levels smaller than 10^{-10}. These results indicate that the CART methodology has captured the structure of a real RNA code, rather than a random phenomenon. Thus, it is an appropriate methodology for analyzing the attributes of the RNA code.

DISCUSSION

Comparing the ensemble attributes of the RNA code with analogous attributes of the genetic code, it is plausible to infer that the essential difference between their degrees of degeneracy (Table 1) reflects the thousand-fold greater "sequence space" available to the former code. Even without counting possible NWC pairing, the RNA code encompasses 4^8 possibilities compared to only 4^3 in the genetic code. For both domains, apparently, the RNA code degeneracy conveys an evolutionary signal; for Archaea it also enfolds an environmental indicator. Prominently, in both domains, the largest degree of degeneracy is for arginine. However, by restricting our vista exclusively to the acceptor stem, the number of sequences that emerge as compatible with aminoacylation by the different amino acids may be artifactually too large. By including nucleotides outside the acceptor stem, that is, by considering the "generalized operational RNA code," the number of determinants for an amino acid may be significantly lowered. It is well known, that 20A in the D-loop is the dominant aminoacylation identity element for this amino acid by ArgRS (Cavarelli et al. 1998; Hendrickson 2001; Shimada et al. 2001). We surmise that for arginine, the exceptionally large degree of RNA code degeneracy indicates a weakening of the selective constraints on the tRNA acceptor stem identity elements. The experimental results lend confirmation of the CART results. In the course of evolution the RNA code for arginine was superseded by the generalized RNA code, resulting in an ultimate reduction of degeneracy for arginine. The archaeal RNA code degeneracy enfolds an ecological factor as well. Adaptation to growth at high temperatures of thermophiles and hyperthermophiles is manifested, inter alia, by distinguishable patterns of amino acid composition, the most noticeable being a twofold decrease in the frequency of glutamine in comparison to that in mesophiles (Singer and Hickey 2003). In our data set of 21 species of Archaea, nine are hyperthermophiles, and five are thermophiles (Supplemental Table S2). In contrast, only six out of 102 bacterial species are thermophiles, 94 are mesophiles, and two are psychrophiles (Supplemental Table S3). This

TABLE 3. Ambiguities in the RNA codes of 14 species of Bacteria

Leaf number[a]	Species	Phylum	Amino acid	tRNA acceptor stem				Anticodon	aaRS class
				N73	N1–N7	N72–N66	(y) WC; (n) NWC		
13	*Thermus thermophilus*[b]	Deinococcus-Thermus	Met	A	GooooGU	oCCoGoo	yyyyyyy	CAU	1
			Lys2	A	GooooGU	oCCoGoo	yyyyyyy	UUU	2
25	*Thermotoga maritime*[c]	Thermotogae	Ile	A	GooCUCG	oCCoAGo	yyyyyyy	GAU	1
			Met	A	GooCUCG	oCCoAGo	yyyyyyy	C_{34}AU	1
37	*Pirellula sp*	Planctomycetacia	Leu	A	GooGoUG	oCCoGAC	nyyyyyy	UAG	1
			Met	A	GooGoUG	oCCoGAC	nyyyyyy	C_{34}AU	1
38	*Nitrosomonas europaea*[c]	b-proteobacteria	Ile	A	GooUoUo	oCCoGAC	yyyyyyy	GAU	1
			Met	A	GooUoUo	oCCoGAC	yyyyyyy	C_{34}AU	1
50	*Rhodopseudomonas palustris*[b]	a-proteobacteria	Val	A	GoooooA	oCCGCoU	yyyyyyy	CAC	1
			Lys2	A	GoooooA	oCCGCoU	yyyyyyy	CUU	2
	Bradyrhizobium japonicum[b]	a-proteobacteria	Val	A	GoooooA	oCCGCoU	yyyyyyy	CAC	1
			Lys2	A	GoooooA	oCCGCoU	yyyyyyy	CUU	2
91	*Campylcobacter jejuni*	e-proteobacteria	Phe	A	GoUUooA	oCAooCU	yyyynyy	GAA	2
			Pro	A	GoUUooA	oCAooCU	yyyynyy	UGG	2
152	*Leifsonia xylixyli*[d]	Actinobacteria	Leu	A	ooCCCCA	CGooGGU	yyyyyyy	UAA	1
			Arg	A	ooCCCCA	CGooGGU	yyyyyyy	CCG	1
278	*Bartonella quintana*[b]	a-proteobacteria	Leu	A	ooCooGA	oGoGUoo	yyyyyyy	CAG	1
			Phe	A	ooCooGA	oGoGUoo	yyyyyyy	CAU	2
	Blochmannia floridanus[b]	g-proteobacteria	Leu	A	ooCooGA	oGoGUoo	yyyyyyy	CAG	1
			Phe	A	ooCooGA	oGoGUoo	yyyyyyy	CAU	2
	Mesorhizobium loti[b]	a-proteobacteria	Leu	A	ooCooGA	oGoGUoo	yyyyyyy	CAG	1
			Phe	A	ooCooGA	oGoGUoo	yyyyyyy	CAU	2
	Pseudomonas syringae[b]	g-proteobacteria	Leu	A	ooCooGA	oGoGUoo	yyyyyyy	CAG	1
			Phe	A	ooCooGA	oGoGUoo	yyyyyyy	CAU	2
382	*Photorhabdus luminescens*[d]	g-proteobacteria	Glu	G	GUCCCCo	CAooooA	yyyyyyy	UUC	1
			Arg	G	GUCCCCo	CAooooA	yyyyyyy	CCU	1
386	*Vibrio vulnificus*[b]	g-proteobacteria	Gln	G	Aoooooo	ooCAoAo	yyyyyyy	UUG	1
			Lys2	G	Aoooooo	ooCAoAo	yyyyyyy	UUU	2
403	*Campylcobacter jejuni*	e-proteobacteria	Asn	G	UooGGoo	oGGoCoA	yyyyyyy	GUU	1
			Asp	G	UooGGoo	oGGoCoA	yyyyyyy	GUC	1

Amino acids are denoted by standard three-letter abbreviations. Nucleotides are denoted by standard one-letter abbreviations. (N73–N66, N1–N7) Nucleotide locations on the tRNA acceptor stem: zeros indicate acceptor stem locations that are occupied by nucleotides that are not part of the bacterial operational RNA code.
[a]See Supplemental Table S4.
[b]Species in which the RNA code ambiguities are resolved by incorporation of synthetases-class information into CART.
[c]Species in which the RNA code ambiguities enciphered by identical tRNA acceptor stem determinants are resolved by post-transcriptional modification of the wobble base C_{34} in the anticodon of $tRNA^{Met}$, with the result that the modified anticodon encodes isoleucine rather than methionine, thereby eliminating the "Met-Ile" RNA code ambiguity in the affected Bacteria.
[d]Species in which the RNA code ambiguities enciphered by identical tRNA acceptor stem determinants are resolved by applying the rules of the generalized operational RNA code.

bias in habitat preference between Archaea and Bacteria helps explain the difference in the degree of degeneracy for glutamine between the two domains, in agreement with the Singer and Hickey (2003) findings. Moreover, it has been demonstrated that the tRNAs of hyperthermophilic Archaea living at temperatures higher than 90°C–95°C harbor exceptionally G-C-rich thermostable stems (including the acceptor), whereas those of thermophilic and mesophilic prokaryotes do not (Marck and Grosjean 2002). It is plausible that part of the degeneracy in the archaeal RNA code is due to adaptations to a hyperthermophilic habitat.

The taxon-specific alternative RNA codes are in principle comparable to the alternative genetic codes (Elzanowski and Ostell 2007). However, code ambiguities have only scarce analogies in the genetic code. Some of the ambiguities detected in this work may be artifactually due to a variety of factors. First, our data may contain sequencing errors. Second, genome annotation is imperfect, and our data may contain pseudogenes erroneously annotated as functional tRNA-specifying genes. Third, by restricting our vista to the acceptor stem and ignoring the generalized operational RNA code for amino acids, we may increase the incidence of multiplicities in the RNA code. Fourth, post-transcriptional modifications may reduce or even eliminate the incidence of particular multiplicities in the RNA code. Finally, in vivo the full complement of synthetases is

employed, which necessitates inclusion of synthetase-class information into the CART input. All these previous factors may inflate the degree of ambiguity. We note, however, that the inclusion of anti-determinants in the data analysis may reduce the incidence of code ambiguity.

Let us now address these issues. Sequencing errors are relatively infrequent and may account for only a few multiplicities. In particular, we removed all sequences with more than two NWC pairs as possibly due to sequencing errors. tRNA pseudogenes are frequently found in vertebrate genomes (Schmitz et al. 2004), but apparently there are only very few such sequences in the genomes of prokaryotes (Giroux and Cedergren 1988). Thus, while erroneous annotation of tRNA pseudogenes as functional genes cannot be completely excluded, we consider it unlikely that such incidences significantly affect our results.

To account for major aminoacylation determinants residing outside the acceptor stem, we applied known rules of the generalized RNA code to the ambiguities found in the terminal leaves of the classification and regression trees. Nucleotide 20A in the D-loop is the principal determinant for aminoacylation by ArgRS (e.g., Shimada et al. 2001). Applying this rule resolves Lys1, Lys2, Asp/Arg ambiguities encrypted in 37 of the 879 archaeal acceptor stems (Supplemental Table S5) and Leu, Glu/Arg ambiguities (Table 3) encrypted in four of the 4011 bacterial tRNAs. The effect of the long variable arm of $tRNA^{Ser}$ as the major aminoacylation determinant for SerRs is to resolve Pro, Asn, Asp/Ser ambiguities found in six archaeal tRNA acceptor stems (Supplemental Table S5), but has no effect on the degree of bacterial RNA code ambiguity. Whether other major determinants outside the acceptor stem can resolve some of the residual 94 archaeal and 26 bacterial code ambiguities in the generalized RNA code remains to be explored experimentally.

Muramatsu et al. (1988) determined that the anticodon of the *Escherichia coli*'s $tRNA^{Ile2}$ is not AUA but CAU (the anticodon for methionine) with the wobble position C34 post-transcriptionally modified to lysidine (k^2C). In Bacteria the alteration converts the codon specificity of a tRNA with the CAU anticodon from AUG (Met) to AUA (Ile) and the amino acid specificity from methionine to isoleucine, thereby preventing the misreading of AUG as Ile and AUA as Met (Marck and Grosjean 2002; Grosjean and Bjork 2004). Another type of modified C34 that has the same properties as lysidine but is chemically distinct from it has been identified in the anticodon of fully mature $tRNA^{Ile}$ of the archaeon *Haloferax volcanii*. Various other types of as-yet-uncharacterized modified C34 might account for the switching property of a tRNA to decode one codon into another, for example, AUG for Met to AUA for Ile (Grosjean and Bjork 2004). Thus, for particular Bacteria and Archaea in our data set, the in silico Ile/Met ambiguity listed in Table 3 and Supplemental Table S5 may be entirely resolved in vivo by post-transcriptional modification of the wobble base C34 of $tRNA^{Met}$, thereby reducing by four (out of 26) and 12 (out of 94) the ambiguities in the bacterial and archaeal generalized RNA codes, respectively.

To investigate the effects of synthetase-class information on the RNA code and its ensemble attributes, the CART algorithm was run on tRNAs apportioned according to the class of cognate synthetases acylating them (Supplemental Tables S2 and S3). Conceptually, any structural feature in a tRNA can play the role of a blocking element in the recognition of a noncognate synthetase (Giegé et al. 1993). Some researchers believe that each tRNA contains anti-determinants against inappropriate synthetases (e.g., Giegé et al. 1998). Yet, compared to the well-explored field of identity determinants, anti-determinants have been sparsely investigated (Fender et al. 2004), and none of the recorded examples relate to the Archaea and Bacteria listed as encoding ambiguous generalized RNA codes (Table 3; Supplemental Table S5). This does not preclude the possibility that further experimental probing might reveal anti-determinants encoded by particular prokaryotes in some of their tRNA molecules as effectively resolving specific code ambiguities found in our study.

Partitioning the data according to synthetase class yielded the two archaeal trees in Figure 2, A and B, and the two bacterial trees in Supplemental Figures S1 and S2. In both domains they feature ~25% fewer paths than the trees generated by the algorithm without inputting synthetase-class information. Noteworthy, there are roughly three times as many terminal nodes in the trees generated from tRNA stems acylated by class 1 synthetases compared to those generated from tRNA stems acylated by class 2 synthetases: 42 versus 16 in the archaeal data set and 251 versus 78 in the bacterial data set.

The relative ranks of degrees of degeneracy in the RNA codes were essentially found to be insensitive to the mode of running the algorithm. The reduction of ~25% of the number of rules in the trees leads to corresponding reduction of the degrees of degeneracy in the RNA codes for essentially all amino acids.

The addition of information concerning synthetase class significantly reduced the incidence of multiplicities in the trees. For example, there are 38 fewer acceptor stems featuring ambiguities in the generalized archaeal RNA code generated by the algorithm utilizing the synthetases class information, compared to ambiguities found by the algorithm processing only the alphabetic information encoded in the acceptor stems (Supplemental Table S5). The resolved ambiguities restored uniqueness to acceptor stem determinant sequences specifying 10 amino acids. The 44 residual ambiguities in the generalized archaeal RNA code of the nine affected archaeons are shown in Table 4. In the bacterial data set the resolved ambiguities restored uniqueness to 16 acceptor stem determinant sequences specifying six amino acids; the residual six acceptor stems harboring ambiguities and the affected two bacteria are shown in

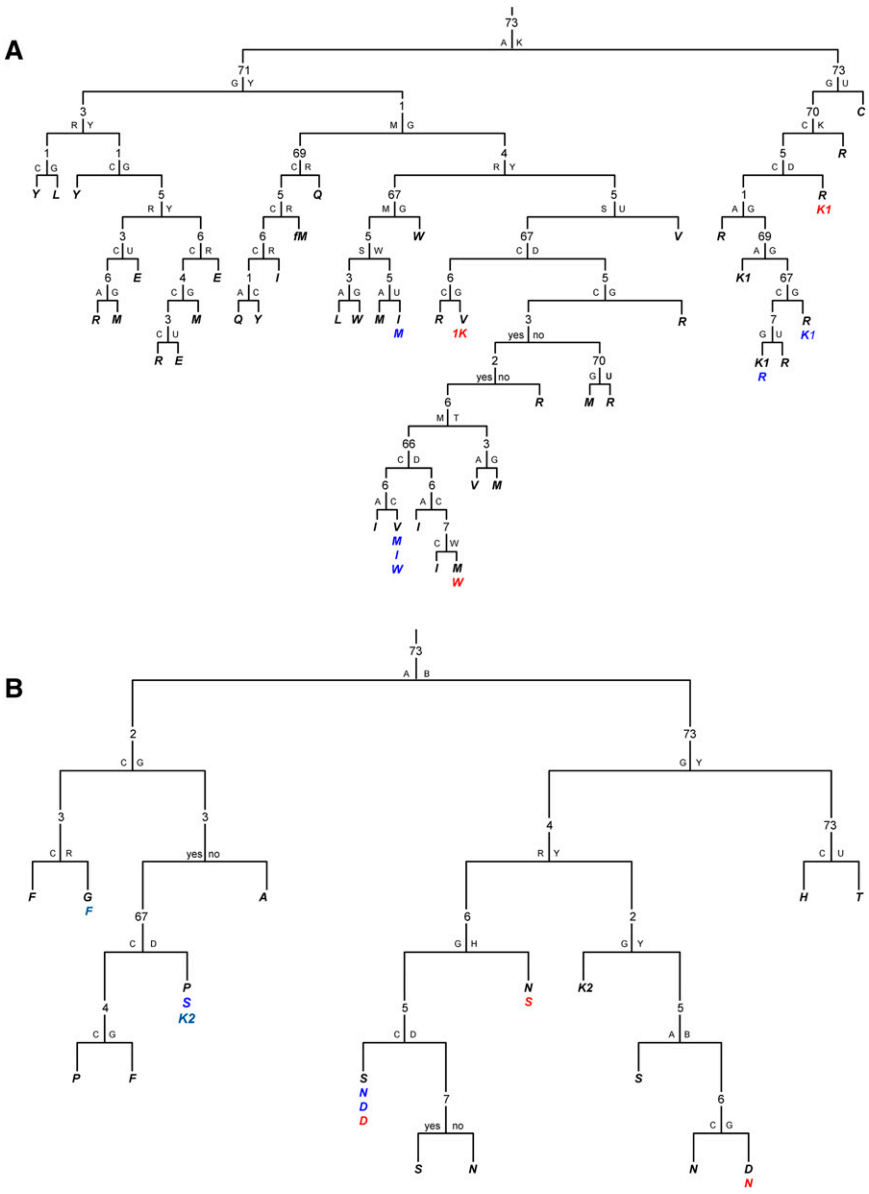

FIGURE 2. Classification and regression trees of domain Archaea after incorporating synthetase-class information. (*A*) The 481 tRNAs acylated by class 1 synthetases generate 42 paths in CART. (*B*) The 398 tRNAs acylated by class 2 synthetases generate 16 paths in CART. For abbreviations and color schemes, see legend to Figure 1.

Table 3. Noteworthy, in both domains, the resolved ambiguities were only those enciphered in acceptor stem determinants of tRNAs acylated by cognate synthetases of opposite class. Conversely, ambiguities residing in acceptor stem determinants of tRNAs acylated by synthetases of the same class were not resolved. A similar pattern pertains for taxon-specific alternative RNA code variants (data not shown). To elucidate the role of class distinction of the synthetases and the significance of the observed pattern of resolved and retained ambiguities in the generalized RNA code, we advance a novel, cryptographic interpretation. The appropriateness of invoking cryptography for elucidating attributes of a biological code is based on its explanatory power. By analogy with the approach of the well-known RSA cryptosystem (Rivest et al. 1978), we denote E for encryption, M for amino acids, C for the enciphered RNA code specifying amino acids by determinants in the tRNA acceptor stem, D for decryption rules embodied in the synthetases, stipulating their class related docking on the 5′ and 3′ branches of the acceptor stem, acylating at the 2′-OH and 3′-OH of the A76 terminus, respectively:

$$E(M) \equiv C \text{ and } D(C) \equiv M.$$

The cryptographic interpretation of our findings is that the synthetase-class distinction is a necessary part of the essential decryption rules for resolving the pairwise RNA code ambiguities enciphered in identical acceptor stem determinants of tRNAs acylated by opposite classes of cognate enzymes. For example, if one of the pair of synthetases recognizing the determinants on one of the tRNA acceptor stem branches acylates the tRNA (harboring ambiguity) with its cognate amino acid, the synthetase of the other class attempting to interact with the same branch will encounter anti-determinants and fail to misacylate the tRNA with the inappropriate amino acid; yet, when interacting with the opposite branch of the same acceptor stem, it will accomplish acylation. By their class division, synthetases are structurally adapted to resolve a specific subset of RNA code ambiguities, thereby restoring uniqueness to amenable determinant sequences. The complementary subset of ambiguities, enciphered in identical acceptor stem determinants of tRNAs amenable to acylation only by synthetases of the same class, is forestalled from affecting the fidelity of protein synthesis by editing reactions occurring at hydrolytic sites embedded in specialized domains of the synthetases (e.g., Nureki et al. 1998; Beebe et al. 2003).

An evolutionary scenario

We conjecture that the output of the CART algorithm when run without information pertaining to synthetase class might be capturing a primordial setting. It is thought that then the hairpin loops of the ancestral tRNA molecules

TABLE 4. Residual ambiguities in the generalized archaeal operational RNA code

Species	Amino acid	tRNA acceptor stem			Anticodon	aaRS class
		N73	N1–N7	N72–N66		
N. equitans	Ile	A	GooCCCo	oCoooGC	GAU	1
	Ile	A	GooCCCo	oCoooGC	UAU	1
	Val	A	GooCCCo	oCoooGC	GAC	1
	Val	A	GooCCCo	oCoooGC	UAC	1
	Val	A	GooCCCo	oCoooGC	CAC	1
S. tokodaii	Ile	A	GooCCCo	oCoooGC	GAU	1
	Val	A	GooCCCo	oCoooGC	GAC	1
	Val	A	GooCCCo	oCoooGC	UAC	1
	Val	A	GooCCCo	oCoooGC	CAC	1
S. solfataricus	Ile	A	GooCCCo	oCoooGC	GAU	1
	Trp	A	GooCCCo	oCoooGC	CCA	1
	Val	A	GooCCCo	oCoooGC	GAC	1
	Val	A	GooCCCo	oCoooGC	UAC	1
	Val	A	GooCCCo	oCoooGC	CAC	1
A. pernix	Ile	A	GooCCCo	oCoooGC	GAU	1
	Trp	A	GooCCCo	oCoooGC	CCA	1
	Val	A	GooCCCo	oCoooGC	GAC	1
	Val	A	GooCCCo	oCoooGC	UAC	1
	Val	A	GooCCCo	oCoooGC	CAC	1
P. horikoshii	Ile	A	GooCCCo	oCoooGC	GAU	1
	Val	A	GooCCCo	oCoooGC	GAC	1
	Val	A	GooCCCo	oCoooGC	UAC	1
	Val	A	GooCCCo	oCoooGC	CAC	1
P. abyssi	Ile	A	GooCCCo	oCoooGC	GAU	1
	Val	A	GooCCCo	oCoooGC	GAC	1
	Val	A	GooCCCo	oCoooGC	UAC	1
	Val	A	GooCCCo	oCoooGC	CAC	1
P. aerophilum	Ile	A	GooCCCo	oCoooGC	GAU	1
	Val	A	GooCCCo	oCoooGC	GAC	1
	Val	A	GooCCCo	oCoooGC	UAC	1
	Val	A	GooCCCo	oCoooGC	CAC	1
P. furiosus	Ile	A	GooCCCo	oCoooGC	GAU	1
	Val	A	GooCCCo	oCoooGC	GAC	1
	Val	A	GooCCCo	oCoooGC	UAC	1
	Val	A	GooCCCo	oCoooGC	CAC	1
M. kandleri	Ile	A	GooCCCo	oCoooGC	GAU	1
	Trp	A	GooCCCo	oCoooGC	CCA	1
	Val	A	GooCCCo	oCoooGC	UAC	1
S. tokodaii	Asn	G	oooGCGo	ooooooo	GUU	2
	Asp	G	oooGCGo	ooooooo	GUC	2
P. aerophilum	Gly	A	oCGoooo	ooooooo	UCC	2
	Gly	A	oCGoooo	ooooooo	CCC	2
	Gly	A	oCGoooo	ooooooo	GCC	2
	Phe	A	oCGoooo	ooooooo	GAA	2

Amino acids are denoted by standard three-letter abbreviations. Nucleotides are denoted by standard one-letter abbreviations. Zeros indicate acceptor stem locations that are occupied by nucleotides that are not part of the archaeal operational RNA code.

were in the process of elongating by addition of base pairs (Di Giulio 1995; Ribas de Pouplana and Schimmel 2001b). The addition of base pairs increased the information encoded in the nascent tRNA acceptor stems, expanding thereby the associated RNA code. That, in turn, provided the platform for accommodating a growing set of primary amino acids in the pre-DNA World. Concurrently, the code ambiguities increased exponentially due to accumulation of identical tRNA acceptor stems related to an increasing variety of amino acids. It has been proposed that proteins may have pre-dated DNA (Dayson 1985; Freeland et al. 1999), among them, possibly, early synthetases. These presumably only would have been capable of interacting with the acceptor stem of present-day tRNA. Therefore, recognition elements responsible for distinguishing among tRNAs would have had to reside only in the acceptor stem (Rould et al. 1989). In evolutionary terms, our findings may be interpreted as indicating that in the pre-DNA World among the nascent proteins interacting with nascent tRNAs, two, with structures enabling containment of the effects of RNA code ambiguities had a selective advantage over proteins lacking such capabilities. Thus, ambiguities in the ancient RNA code may have been pivotal for the fixation of these enzymes in the genomes of ancestral prokaryotes, corroborating Schimmel and Ribas de Pouplana (2001), who surmised that interactions between the catalytic domains of the synthetases with tRNA acceptor stems may have formed the basis for their two classes. Protocells that acquired and fixated synthetases in their genomes attained a crucial advantage over others that failed to do so. The class division of synthetases was (and continues to be) instrumental for sustaining the functional role of the RNA code as the link between enzymes, amino acids, and tRNAs.

Conclusion

In this study we took advantage of the availability of close to 5000 unmodified tRNA sequences belonging to 21 Archaea and 102 Bacteria with completely sequenced genomes in order to investigate with the aid of data-mining statistical tools the ensemble attributes of the prokaryotic RNA codes and find out whether some of these attributes may have been putative agents for the enigmatic origin of the two distinct

classes of synthetases in the genomes of ancestral prokaryotes.

By using a CART statistical methodology, RNA code degeneracy, taxon-specific alternative codes, and codes with ambiguity were identified as key in silico ensemble attributes of the RNA code.

The degeneracy of RNA codes appears to be essentially different from that of the genetic code, due to the thousand-fold greater "sequence space" available (without counting NWC pairings). For both domains it seems to convey also an evolutionary signal: the exceptionally large degree of arginine's RNA code degeneracy (Table 1) indicates a weakening of the selective constraints on its tRNA acceptor stem identity elements. In the course of evolution, its RNA code was superseded by the generalized RNA code, the 20A base in the D-loop becoming the dominant aminoacylation identity element (e.g., Shimada et al. 2001), resulting in the reduction of the degeneracy for arginine to the value of "1." The archaeal RNA code degeneracy enfolds an ecological factor as well. In our data set, two-thirds of the Archaea are hyperthermophiles or thermophiles (Supplemental Table S2), and nine-tenths of the Bacteria are mesophiles (Supplemental Table S3). In conformity with the twofold decrease in the frequency of glutamine found to be the most noticeable change in the pattern of amino acid composition of proteins synthesized by thermophiles in comparison with those by mesophiles (Singer and Hickey 2003), the rankings of glutamine's degeneracy in the archaeal and bacterial RNA codes are 4 and 12, respectively (Table 1).

We found in our data set archaeal and bacterial tRNA acceptor stems that are not unique; that is, two or more full acceptor stem sequences (N1–N7, N72–N66, determinants and positions not participating in the RNA code) are related to the same amino acid. The taxon-specific alternative RNA codes (identical acceptor stem determinants encrypting different amino acids in different species) and the RNA codes with ambiguity (identical acceptor stem determinants encrypting more than one amino acid in the same species) originate in this phenomenon. The former are analogous to the well-known alternative genetic codes (e.g., Elzanowski and Ostell 2007). Ambiguities seem to be in a different category because of the scarcity of analogs in the universal code. We found 137 archaeal and 30 bacterial tRNA acceptor stems to harbor in silico ambiguities in the RNA code. Close scrutiny revealed that the collective properties of the whole set of tRNAs and synthetases within extant cells are able to improve considerably the accuracy of the cellular aminoacylation system. In vivo, we expect 93 and 24 of the respective in silico archaeal and bacterial RNA code ambiguities to be resolved. We conjecture that the remainder, 5.0% versus 0.15%, respectively, found in 44 and six among the 879 and 4011 tRNA stems in our study, constitute an irreducible set. The preponderance of intrinsic archaeal compared to bacterial RNA code ambiguities appears to be a previously unreported specific domain distinction. We predict that it will prevail as more and more archaeal and bacterial tRNA acceptor stems become available from completely sequenced genomes.

By resolving specific ambiguities in the RNA code, our study results point to the structural distinction between the two classes of synthetases very effective among the collective properties of the whole set of tRNAs and synthetases within extant cells, which evolved as such to improve the accuracy of the cellular aminoacylation system. The RNA operational code is considered to have been originally highly ambiguous (e.g., Rodin and Rodin 2008). To investigate the possible evolutionary effect of the inception of the class-distinct synthetases on the high incidence of code ambiguities, trees were generated with data partitioned by the class of synthetases acylating cognate tRNAs in the living cell. This resolved in both domains, a particular subset of the code ambiguities without affecting the complementary subset. Making a novel biological use of an analogy with the well-known RSA cryptosystem (Rivest et al. 1978), a plausible cryptographic interpretation of our findings is that the synthetases' class distinction of synthetases is a necessary part of the essential decryption rules for resolving the pairwise RNA code ambiguities enciphered in identical acceptor stem determinants of tRNAs acylated by opposite class of cognate enzymes. By their class division, synthetases are structurally adapted to resolve only one subset of RNA code ambiguities, but not ambiguities from the complementary subset, enciphered in identical acceptor stem determinants of tRNAs acylated by the same class of synthetases. This result corroborates the Schimmel and Ribas de Pouplana (2001) notion that interactions between tRNA acceptor stems and synthetases formed the basis for their distinct two classes. Ambiguities in the ancient RNA code apparently were pivotal for the fixation of these enzymes in the genomes of ancestral prokaryotes; their class division was and continues to be instrumental for sustaining the functional role of the RNA code as the link between enzymes, amino acids, and tRNAs.

MATERIALS AND METHODS

Data

We extracted 4925 unmodified tRNA sequences belonging to 21 Archaea and 102 Bacteria with completely sequenced genomes from public databases (http://www.ncbi.nlm.nih.gov; http://www.uni-bayreuth.de/department/biochemie/trna; http://lowelab.ucsc.edu/GtRNAdb; http://genome.jpi-psf.org/microbal/index.htm). Species names followed TaxBrowser (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/). Only one strain from each species was included in the data set. If two or more copies of a tRNA with the same anticodon were encountered and if their acceptor stems sequences were identical, then a single sequence was used, thus avoiding considerations of frequency. The

correspondence between amino acids and the order of nucleotides in acceptor stems of tRNAs was determined for the 20 primary amino acids. tRNAs for selenocysteine and pyrrolysine were omitted. A tRNA acceptor stem was defined as comprising 15 bases, the "discriminator base" N73, and seven base pairs: N1:N72, N2:N71, N3:N70, N4:N69, N5:N68, N6:N67, and N7:N66, where N stands for any of the four nucleotides. The invariant trinucleotide CCA at the 3′ terminus was omitted from the data set. Also omitted from the data set was the "G0" position in $tRNA^{His}$. The location of non-Watson–Crick (NWC) pairs was noted. Acceptor stems with three or more NWC pairs were omitted as outliers reflecting possible sequencing errors. The final data set consisted of 4011 bacterial and 879 archaeal acceptor stems (Supplemental Tables S2, S3, respectively).

Notations and abbreviations

Although our data are genomic, we used the RNA alphabet. Seventeen archaeons and 11 bacteria in our data set use LysRS1; 4 archaeons and 91 bacteria use LysRS2. The abbreviations Lys1 (or K1) and Lys2 (or K2) are used to indicate lysine ligated to cognate tRNAs by LysRS1 and LysRS2, respectively. The seven variables associated with each acceptor stem were (1) amino acid; (2,3) taxonomic affiliation at the domain (Archaea or Bacteria) and species levels; (4) the anticodon trinucleotide; (5,6) the alphabetic information encoded in the acceptor stem, that is, the 5′ and the 3′ nucleotide sequences, respectively; and (7) the location of NWC pairs. Compilations of tRNA acceptor stems data, sorted by amino acids, species, class of synthetase, and preferred habitat (e.g., hyperthermophile, thermophile, etc.) are presented in Supplemental Tables S2 and S3. Archaea and Bacteria are analyzed separately in accord with experimentally determined differences in the RNA code between the two prokaryotic domains (Xu et al. 2001).

Classification and regression tree (CART) algorithm

Our main tool for studying the association between amino acids and the sequence of nucleotides of the acceptor stems was the CART algorithm (Breiman et al. 1984; Clark and Pregibon 1993; Kim et al. 2004) as implemented in the statistical software S-PLUS (http://www.insightful.com). In this study, the independent variables were the nucleotides at different positions on the acceptor stem or the NWC status of a nucleotide pair, and the dependent variables were the amino acids. CART is a decision algorithm that recursively divides the data in a binary manner according to certain optimality criteria and stops when certain conditions are met. As optimality criterion we used the deviance (a statistical measure of node purity, which is equal to minus twice the log-likelihood of obtaining a partition under a model of random partitions). The CART algorithm is "greedy," that is, at each step, all possible further data splits are considered at every node. Out of all the possible splits, the one chosen to partition the data was the split with the smallest deviance. At each node, a rule is added to the preceding rules on the path. The process yields a binary decision tree that is allowed to grow as much as possible with no limit to the number of rules. Each node in the tree represents a condition in the form "IF K, THEN go to the left; IF NOT K, THEN go to the right, where "K" represents a subset of nucleotides at a position or the NWC status for a pair of nucleotides in the acceptor stem. Combinations of decision rules create paths from the root of the decision tree to the terminal nodes. A position on the acceptor stem may appear more than once on a path. Terminal nodes specify amino acids corresponding to the set of conditions along the path, that is, the sequences of determinants on the acceptor stems of tRNAs that are compatible with particular amino acids. A path on the tree terminates when a node is reached with a specific amino acid in it, so that the deviance is zero, or if the deviance is larger than zero but cannot be decreased by further splitting. By definition, this means that the terminal node contains two or more amino acid types that cannot be further purified. For such a node, the most frequent amino acid, termed "predominant," is assigned to the node; the others, ranked by their relative frequencies, are termed "misclassified" or "complementary," and the node is referred to as harboring multiplicities. Each of the complementary amino acids is associated with a repetition of the same rules that predicted the most frequent one. The "misclassification error rate" is defined as the ratio "(copy number of misclassified amino acids)/(number of acceptor stems)." The success of the CART algorithm is judged by the magnitude of the misclassification error rate on the entire tree. If a position in the acceptor stem does not appear along a path, then any one of the four bases can occupy this position without affecting the identity of the amino acid at the terminal node (verifiable by inserting the four bases sequentially in such a position along the path). Consequently, such positions are not part of the operational RNA code for its ensemble of species encoding the tRNAs in the data set.

Permutation tests

The statistical significance of the set of rules deduced by the CART algorithm was assessed by permutation tests. The correspondence found by the algorithm between the determinants encrypted in specific tRNA acceptor stems and associated amino acids was severed, and the amino acids were randomly assigned to acceptor stems, while maintaining the nucleotide order in the stems intact. Each simulation consisted of 1000 runs. The number of inferred rules and the misclassification error rate on the trees generated by the CART algorithm were compared to the respective means of the simulation.

SUPPLEMENTAL MATERIAL

Supplemental material can be found at http://www.rnajournal.org.

ACKNOWLEDGMENTS

We thank Tal Pupko for comments and suggestions, Tal Galili for help with the data analysis, and Giddy Landan for a great number of favors, large and small. We also thank the manuscript reviewers for many useful criticisms and suggestions that have greatly enhanced its quality and readability. D.G. and D.B. were supported by a Small Grant Award from the University of Huston and by the U.S. National Library of Medicine grant LM010009-01 to D.G. and Giddy Landan. Y.B. was supported in part by a grant from the Israel Science Foundation.

Received May 26, 2009; accepted September 30, 2009.

REFERENCES

Arnez JG, Harris DC, Mitschler A, Rees B, Francklyn CS, Moras D. 1995. Crystal structure of histidyl-tRNA synthetase from *Escherichia coli* complexed with histidyl-adenylate. *EMBO J* **14:** 4143–4155.

Beebe K, Ribas de Pouplana L, Schimmel P. 2003. Elucidation of tRNA-dependent editing by a class II tRNA synthetase and significance for cell viability. *EMBO J* **22:** 668–675.

Beebe K, Merriman E, Ribas De Pouplana L, Schimmel P. 2004. A domain for editing by an archaebacterial tRNA synthetase. *Proc Natl Acad Sci* **101:** 5958–5963.

Beuning PJ, Musier-Forsyth K. 1999. Transfer RNA recognition by amynoacyl-tRNA synthetases. *Biopolymers* **52:** 1–28.

Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. *Classification and regression trees*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Cavarelli J, Moras D. 1993. Recognition of tRNAs by aminoacyl-tRNA synthetases. *FASEB J* **7:** 79–86.

Cavarelli J, Delagoutte B, Erani G, Ganiloff J, Moras D. 1998. L-arginine recognition by yeast argynyl-tRNA synthetase. *EMBO J* **17:** 5438–5448.

Choi H, Otten S, Schneider J, McClain WH. 2002. Genetic perturbations of RNA reveal structure-based recognition in protein–RNA interaction. *J Mol Biol* **324:** 573–576.

Clark LA, Pregibon D. 1993. Tree-based models. In *Statistical models in S* (ed. JM Chambers, TJ Hastie), pp. 393–395. Chapman & Hall, London.

Dayson F. 1985. *Infinite in all directions*, pp. 54–73. Harper & Row, New York.

Di Giulio M. 1995. Was it an ancient gene codifying for a hairpin RNA that, by means of direct duplication, gave rise to the primitive tRNA molecule? *J Theor Biol* **177:** 95–101.

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

Elzanowski A, Ostell J. 2007. *The genetic codes*. www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi.

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 motif. *Nature* **347:** 203–206.

Fender A, Sissler M, Florentz C, Giege R. 2004. Functional idiosyncrasies of tRNA isoacceptors in cognate and noncognate aminoacylation systems. *Biochimie* **86:** 21–29.

Fersht AR. 1998. Sieves in sequence. *Science* **280:** 541–543.

Francklyn C, Schimmel P. 1989. Aminoacylation of RNA minihelices with alanine. *Nature* **337:** 478–481.

Francklyn C, Schimmel P. 1990. Enzymatic aminoacylation of an eight-base-pair microhelix with histidine. *Proc Natl Acad Sci* **87:** 8655–8659.

Freeland SJ, Knight RD, Landweger FL. 1999. Do proteins predate DNA? *Science* **286:** 690–692.

Fukai S, Nureki O, Sekine S, Shimada A, Tao J, Vassylyev DG, Yokoyama S. 2000. Structural basis for double-sieve discrimination of L-valine from L-isoleucine and L-threonine by the complex of $tRNA^{Val}$ and valyl-tRNA synthetase. *Cell* **103:** 793–803.

Giegé R, Kern D, Ebel J-P, Grosjean H, De Henau S, Chantrenne H. 1974. Incorrect aminoacylation involving tRNAs or Vanyl-tRNA synthetase from *Bacillus stearothermophilus*. *Eur J Biochem* **45:** 351–362.

Giegé R, Puglisi JD, Florentz C. 1993. tRNA structure and aminoacylation efficiency. *Prog Nucleic Acid Res Mol Biol* **45:** 129–206.

Giegé R, Sissler M, Florentz C. 1998. Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res* **26:** 5017–5035.

Giroux S, Cedergren R. 1988. Tandemly repeated tRNA pseudogenes in photobacterium. *Proc Natl Acad Sci* **85:** 9101–9105.

Goldgur Y, Mosyak L, Reshetnikova A, Ankilova V, Latvik O, Khodyreva S, Safro M. 1997. Crystal structure of phenylalanyl-tRNA synthetase from *Thermus thermophilus* complexed with cognate $tRNA^{Phe}$. *Structure* **15:** 59–68.

Graur D, Li W-H. 2000. *Fundamentals of molecular evolution*, 2nd ed., p. 7. Sinauer Associates, Inc., Sunderland, MA.

Grosjean H, Bjork GR. 2004. Enzymatic conversion of cytidine to lysidine in anticodon of bacterial $tRNA^{Ile}$—an alternative way of RNA editing. *Trends Biochem Sci* **29:** 165–168.

Grothers DM, Seno T, Söll DG. 1972. Is there a discriminator site in transfer RNA? *Proc Natl Acad Sci* **69:** 3063–3067.

Hamann GS, Hou YM. 1995. Enzymatic aminoacylation of tRNA acceptor stem helices with cysteine is dependent on a single nucleotide. *Biochemistry* **34:** 6527–6532.

Hendrickson TL. 2001. Recognizing the D-loop of transfer RNAs. *Proc Natl Acad Sci* **98:** 13473–13475.

Ibba M, Morgan S, Curnow AW, Pridmore DR, Vothknecht UC, Gardner W, Lin W, Woese C, Söll D. 1997. A eukyarchaeal lysyl-tRNA synthetase: Resemblance to class-1 synthetases. *Science* **278:** 1119–1122.

Ibba M, Losey HC, Kawarabayasi Y, Kikuchi H, Bunjun S, Söll D. 1999. Substrate recognition by class-I lysyl-tRNA synthetases: A molecular basis for gene displacement. *Proc Natl Acad Sci* **96:** 418–423.

Jones TE, Brown CL, Geslain R, Alexander RW, Ribas de Pouplana L. 2008. An operational RNA code for faithful assignment of AUG triplets to methionine. *Mol Cell* **29:** 401–407.

Jukes TH, Osawa S. 1997. Point counter point. Further comments on codon reassignment. *J Mol Evol* **45:** 1–3.

Kim H, Guess FM, Young M. 2004. *Using data mining tools of decision trees in reliability applications*, pp. 1–20. University of Tennessee Monograph. http://bus.utk.edu/soms/Information/forms/2004-02.pdf.

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

Lovato MA, Chihade JW, Schimmel P. 2001. Translocation within the acceptor helix of a major tRNA identity determinant. *EMBO J* **20:** 4846–4853.

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

Martinis SA, Schimmel P. 1993. Microhelix aminoacylation by a class I tRNA synthetase. Nonconserved base pairs required for specificity. *J Biol Chem* **268:** 6069–6072.

Martinis SA, Schimmel P. 1995. Small RNA oligonucleotide substrates for specific aminoacylations. In *tRNA structure, biosynthesis and function* (ed. D Söll, UL RajBhandary), pp. 349–370. ASM Press, Washington DC.

McClain WH. 1993. Rules that govern tRNA identity in protein synthesis. *J Mol Evol* **234:** 257–280.

McClain WH, Schneider J, Bhattacharya S, Gabriel J. 1998. The importance of tRNA backbone mediated interactions with synthetases for aminoacylation. *Proc Natl Acad Sci* **95:** 460–465.

Möller W, Janssen GMC. 1990. Transfer RNAs for primordial amino acids contain remnants of a primitive code at positions 3 to 5. *Biochimie* **72:** 361–368.

Muramatsu T, Yokoyama S, Horie N, Matsuda A, Ueda T, Yamaizumi Z, Kuchio Y, Nishimura S, Miyazawa T. 1988. A novel lysine-substituted nucleoside in the first position of the anticodone of minor isoleucine tRNA from *Escherichia coli*. *J Biol Chem* **263:** 9261–9267.

Musier-Forsyth K, Schimmel P. 1999. Atomic determinants for aminoacylation of RNA minihelices and relationship to genetic code. *Acc Chem Res* **32:** 368–375.

Nair S, Ribas de Pouplana L, Houman F, Avruch A, Shen H, Schimmel P. 1997. Species-specific tRNA recognition in relation to tRNA synthetase contact residues. *J Mol Biol* **269:** 1–9.

Nureki O, Niimi T, Muramatsu T, Kanno H, Kohno T, Florentz C, Giegé R, Yokoyama S. 1994. Molecular recognition of the identity-determinant set of isoleucine transfer RNA from *Escherichia coli*. *J Mol Biol* **236:** 710–724.

Nureki O, Vassylyev DG, Tateno M, Shimada A, Nakama T, Fukai S, Konno M, Hendrickson TL, Schimmel P, Yokoyama S. 1998. Enzyme structure with two catalytic sites for double-sieve selection of substrate. *Science* **280:** 578–582.

Park SJ, Schimmel P. 1988. Evidence for interaction of an aminoacyl transfer RNA synthetase with a region important for the identity of its cognate transfer RNA. *J Biol Chem* **263:** 16527–16530.

Ribas de Pouplana L, Schimmel P. 2001a. Operational RNA code for amino acids in relation to genetic code in evolution. *J Biol Chem* **276:** 6881–6884.

Ribas de Pouplana L, Schimmel P. 2001b. Aminoacyl-tRNA synthetases: Potential markers of genetic code development. *Trends Biochem Sci* **26:** 591–596.

Ribas de Pouplana L, Schimmel P. 2001c. Two classes of tRNA synthetases suggested by sterically compatible docking on tRNA acceptor stem. *Cell* **104:** 191–193.

Rivest RL, Shamir A, Adleman L. 1978. A method for obtaining digital signatures and public-key cryptosystems. *Commun ACM* **21:** 120–126.

Rodin SN, Rodin AS. 2008. On the origin of the genetic code: Signatures of its primordial complementarity in tRNAs and aminoacyl-tRNA synthetases. *Heredity* **100:** 341–355.

Rould MA, Perona JJ, Söll D, Steitz TA. 1989. Structure of *E. coli* glutaminyl-tRNA synthetase complexed with $tRNA^{Gln}$ and ATP at 2.8 Å resolution. *Science* **189:** 1135–1142.

Saks ME, Sampson JR. 1996. Variant minihelix RNAs reveal sequence-specific recognition of the helical $tRNA^{Ser}$ acceptor stem by *E.coli* seryl-tRNA synthetase. *EMBO J* **15:** 2843–2849.

Santos MA, Cheesman C, Costa V, Moradas-Ferreira P, Tuite MF. 1999. Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in *Candida spp*. *Mol Microbiol* **31:** 937–947.

Santos MA, Moura G, Massey SE, Tuite MF. 2004. Driving change: The evolution of alternative genetic codes. *Trends Genet* **20:** 95–102.

Schimmel P. 1995. An operational RNA code for amino acids and variations of critical nucleotide sequences in evolution. *J Mol Evol* **40:** 531–536.

Schimmel P, Ribas de Pouplana L. 1995. Transfer RNA: From minihelices to genetic code. *Cell* **81:** 983–986.

Schimmel P, Ribas de Pouplana L. 2001. Formation of two classes of tRNA synthases in relation to editing functions and genetic code. In *Cold Spring Harbor Symposium on Quantitative Biology: The Ribosome*, Vol. 66, pp. 16–166. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. 16–166.

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* **90:** 8763–8768.

Schmitz J, Churakov G, Zischler H, Brosius J. 2004. A novel class of mammalian-specific tailless retropseudogenes. *Genome Res* **14:** 1911–1915.

Schultz DW, Yarus M. 1996. On malleability in the genetic code. *J Mol Evol* **42:** 597–601.

Sherman JM, Rogers MJ, Söll D. 1992. Competition of aminoacyl-tRNA synthetases for tRNA ensures the accuracy of aminoacylation. *Nucleic Acids Res* **20:** 1547–1552.

Shimada A, Nureki O, Goto M, Takahashi S, Yokoyama S. 2001. Structural and mutational studies of the recognition of the arginine tRNA-specific major identity element, A20, by arginyl-tRNA synthetase. *Proc Natl Acad Sci* **98:** 13537–13542.

Singer GAC, Hickey DA. 2003. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* **317:** 39–47.

Stehlin C, Burke B, Yang F, Liu H, Shiba K, Musier-Forsyth K. 1998. Species-specific differences in the operational RNA code for aminoacylation of $tRNA^{Pro}$. *Biochemistry* **37:** 8605–8613.

Suzuki T, Ueda T, Watanabe K. 1997. The 'polysemous' codon—a codon with multiple amino acid assignments caused by dual specificity of tRNA identity. *EMBO J* **16:** 1122–1134.

Weygand-Durasevic I, Gruic-Sovuij I, Rocak S, Landeka I. 2002. The accuracy of seryl-tRNA synthesis. *Food Technol Biotechnol* **40:** 247–253.

Wolfson AD, LaRiviere FJ, Pleiss JA, Dale T, Asahara H, Uhlenbeck OC. 2001. tRNA conformity. *Cold Spring Harb Symp Quant Biol* **66:** 185–193.

Xu F, Chen X, Xin L, Chen L, Jin Y, Wang D. 2001. Species-specific differences in the operational RNA code for aminoacylation of $tRNA^{Trp}$. *Nucleic Acids Res* **29:** 4125–4133.

Yarus M, Schultz DW. 1997. Point counter point. *J Mol Evol* **45:** 3–6.