Imprints of the genetic code in the ribosome

David B. F. Johnson and Lei Wang[1]

The Jack H. Skirball Center for Chemical Biology and Proteomics, The Salk Institute for Biological Studies, La Jolla, CA 92037

Edited* by Peter G. Schultz, The Scripps Research Institute, La Jolla, CA, and approved March 18, 2010 (received for review January 19, 2010)

The establishment of the genetic code remains elusive nearly five decades after the code was elucidated. The stereochemical hypothesis postulates that the code developed from interactions between nucleotides and amino acids, yet supporting evidence in a biological context is lacking. We show here that anticodons are selectively enriched near their respective amino acids in the ribosome, and that such enrichment is significantly correlated with the canonical code over random codes. Ribosomal anticodon-amino acid enrichment further reveals that specific codons were reassigned during code evolution, and that the code evolved through a two-stage transition from ancient amino acids without anticodon interaction to newer additions with anticodon interaction. The ribosome thus serves as a molecular fossil, preserving biological evidence that anticodon-amino acid interactions shaped the evolution of the genetic code.

evolution of the genetic code | stereochemical hypothesis | anticodon-amino acid association | codon reassignment | RNA-protein interaction

The origin and evolution of the genetic code is a critical transition in the evolution of all modern organisms (1). Understanding why the genetic code evolved to its modern form is as important, if not more so, as knowing the code itself (2), as it is central to understanding major evolutionary breakthroughs. However, the universality of the code is also its downfall with regard to studying its formation, as no organisms exist containing a primitive or intermediate genetic code for comparison. Although multiple hypotheses have been proposed to explain why codons are selectively assigned to specific amino acids (3, 4), empirical data are extremely rare and difficult to obtain (5–8), leaving many theories in the realm of conjecture. One theory addressing this challenging question is the stereochemical hypothesis, which postulates that the genetic code developed from interactions between anticodon- or codon-containing polynucleotides and their respective amino acids (5, 9). This theory is supported by RNA aptamer experiments, in which RNA molecules evolved to bind specific amino acids in vitro are enriched with anticodon and codon elements for the amino acid (6–8, 10). Codons for arginine have also been found to confer binding specificity for arginine in the self-splicing group I introns (11). Nonetheless, the biological relevance of these aptamers and introns to genetic code evolution is unknown, and no further in vivo data exist to support this hypothesis.

If chemical or physical interactions between nucleotides and amino acids did influence the evolution of the genetic code, relics of this evolution may be present in modern cells. In particular, we may uncover such imprints within RNA-binding proteins and RNA-protein interactions. The ribosome presents an excellent model for the study of these potential interactions, as it is a large ribonucleoprotein complex with some 50 proteins interacting extensively with the ribosomal RNAs (rRNAs) for stability and structure (Fig. 1*A*) (12). In addition, the ribosome emerged from an early evolution stage of life to establish the translation of the genetic code before the last universal common ancestor (LUCA) (13), and thus is more likely to preserve relics of the underlying force driven the formation of the code. A comprehensive analysis of ribosome structural data will help to reveal if such interactions survive in the ribosome and if they correlate with the canonical genetic code. The correlation, if established, would provide empirical in vivo data from modern organisms for understanding the origin and evolution of the genetic code.

Results

Codons or Anticodons for Select Amino Acids Are Enriched Near Their Respective Amino Acids in Ribosome Structures. We examined whether codons (or anticodons) assigned to each amino acid by the canonical genetic code were enriched near the respective amino acid in the ribosome. Structures of ribosomes from four species (one archaebacterium and three eubacteria) (14–17) were analyzed. Many amino acids were found to contact their codon (or anticodon) sequences in rRNAs (Fig. 1*B*). A codon (or anticodon) enrichment value for an amino acid was calculated from the probability of finding a block of triplet codon (or anticodon) sequences within 5 Å of the amino acid relative to other 19 amino acids in the ribosomal structures (Fig. 1*C*) (*Methods*). To determine if a global trend of codon or anticodon enrichment was significant, independent statistical tests of each enriched amino acid were combined using Fisher's method for each analysis. Using a statistical cutoff of $P < 0.05$, 11 amino acids were found to be significantly enriched with anticodons ($P = 0.039$) and 8 with codons ($P = 0.045$) (Fig. 1 *D* and *E*). Two amino acids (Leu and Ile) are present in both categories. These results demonstrate that, for a subset of amino acids, codons or anticodons are enriched near their respective amino acids in the ribosome.

Ribosomal Anticodon-Amino Acid Enrichment Is Correlated with the Canonical Genetic Code. Is the codon (or anticodon)-amino acid enrichment observed in the ribosome correlated with the canonical genetic code, or can similar results be obtained with random codes? We applied a series of Monte Carlo simulations to determine if the amino acid-codon assignment in the canonical genetic code is better at explaining the enrichment than the assignments in random codes. We first analyzed the probability of finding a better code using only those amino acids in Fig. 1 *D* and *E*. One million random codes were generated. The average enrichment values for both the anticodon- and codon-enriched amino acids were determined using the ribosome structural data for each code, and were compared to those determined from the canonical code (1.23 for anticodon-enriched and 1.17 for codon-enriched amino acids; *Methods*). The distribution of the average enrichment values in the random codes is shown in Fig. 2*A* for codons (*Left*) and anticodons (*Right*). For comparison, the average enrichment value for the canonical code is denoted by an arrow in each of the histograms. For the 11 amino acids enriched with anticodons and 8 enriched with codons, 99.9555% and 99.9716% of random codes have lower average enrichment than the canonical code, respectively.

The next challenge in these analyses is to demonstrate that the enrichments in the specific subsets are not a result of random chance. To investigate whether correlations can also be found in random codes with different subsets of amino acids enriched with codons (or anticodons) in the ribosome, we next performed a global correlation analysis and an optimal subset correlation analysis. In

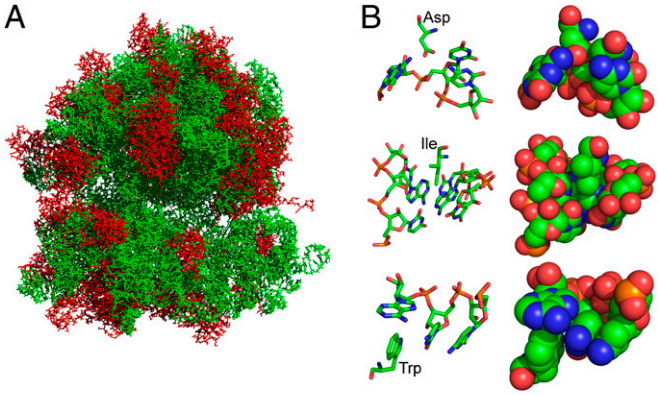
Author contributions: D.B.F.J. and L.W. designed research; D.B.F.J. performed research; D.B.F.J. and L.W. analyzed data; and D.B.F.J. and L.W. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]To whom correspondence should be addressed. E-mail: lwang@salk.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/1000704107/DCSupplemental.



C

Amino Acid	E_x (Codons)	E_x (Anticodons)
Ala	0.912	0.951
Arg (CGN)	1.034	1.014
Arg (AGR)	0.929	0.998
Asn	0.796	0.987
Asp	0.989	1.320
Gln	0.946	1.234
Glu	1.149	0.596
Gly	1.058	1.006
His	0.904	1.078
Ile	1.437	1.283
Leu (UUR)	1.403	1.119
Leu (CUN)	1.059	1.050
Lys	1.033	0.955
Met	0.731	1.152
Phe	0.862	1.384
Pro	0.980	0.993
Ser (UCN)	0.954	0.988
Ser (AGY)	0.984	1.042
Thr	1.156	0.923
Trp	0.667	1.640
Tyr	1.130	1.274
Val	0.916	0.922

D

Amino Acid	*P*-value	-2ln(*P*)
Ile	0.0225	7.588
Leu (UUR)	0.0652	5.461
Thr	0.0816	5.012
Glu	0.2409	2.847
Gly	0.2998	2.409
Arg (CGN)	0.4420	1.633
Lys	0.6437	0.881
Tyr	0.6513	0.858
Sum (χ^2)		26.69
P-value		0.0451

E

Amino Acid	*P*-value	-2ln(*P*)
Phe	0.0118	8.879
Trp	0.1052	4.504
Asp	0.1118	4.382
Ile	0.1169	4.293
Gln	0.1182	3.852
Tyr	0.1457	1.462
Leu (UUR)	0.4814	1.196
His	0.5498	1.046
Leu (CUN)	0.5962	1.046
Arg (CGN)	0.7596	0.055
Met	0.7834	0.488
Sum (χ^2)		34.92
P-value		0.0395

Fig. 1. For a subset of amino acids, codons or anticodons are enriched near their respective amino acids in ribosome structures. (*A*) The structure of *Thermus thermophilus* ribosome with proteins highlighted in red and rRNAs in green [adapted from PDB 2J00 and 2J01, (14)]. (*B*) Examples of amino acid contacting its anticodon rRNA shown in the stick (*Left*) and spherical (*Right*) representation. (*Top*) Asp with GUC; (*Middle*) Ile with 2 UAU; (*Bottom*) Trp with CCA [*Top* and *Middle* representations are adapted from *Escherichia coli* ribosome, PDB 2QBA (15); *Bottom* representation is adapted from *Haloarcula marismortui* ribosome, PDB 1VQO (17)]. (*C*) Relative enrichment values of codons and anticodons near their respective amino acids in ribosome structures from four different species. Amino acids with six codons were split into a four-codon block and a two-codon block, which were calculated independently. (*D*) Eight amino acids showed significant enrichment of codon-containing rRNA. (*E*) Eleven amino acids showed significant enrichment of anticodon-containing rRNA. See *Methods* for a detailed description of the formulas used to calculate the enrichment values. χ^2 analysis was performed for each amino acid using the enrichment values and combined using Fisher's method.

the global correlation analysis, all amino acids (excluding Cys because of too few Cys available at the surface of ribosomal proteins to interact with rRNA; *Methods*) were taken into consideration to calculate the average enrichment values based on the ribosomal structural data. For the enrichment of anticodons, 99.0225% of randomized codes produced a lower average enrichment than the canonical code (Fig. 2*B*, *Right*). In contrast, the enrichment of codons no longer showed correlation with the canonical code, with only 54.5447% of random codes producing a lower average enrichment value (Fig. 2*B*, *Left*). In the optimal subset correlation analysis, we used the subset of 8 amino acids most enriched with their codons and of 11 amino acids with their anticodons in the ribosome for each randomly generated code. The identity of amino acids in the subset thus varied with the code. Distributions for these analyses were plotted and compared to the canonical code for codons and anticodons, as before (Fig. 2*C*). Results are similar to the global correlation analysis, with 71.0076% and 95.1925% of random codes having lower enrichment of codons and anticodons, respectively. These two analyses suggest that the codon enrichment with the specific subset of amino acids cannot be attributed to more than random chance, as a significantly large portion of random codes generate better results with either an optimal set of eight amino acids or the entire set. The anticodon enrichment of the canonical code proves to be stronger than a significant portion of random codes in all analyses, demonstrating a strong correlation between the ribosomal anticodon-amino acid enrichment and the canonical code that cannot be attributed to random chance alone. Such a correlation suggests that amino acid-anticodons interactions could have contributed to the organization of the canonical genetic code, supporting the stereochemical hypothesis.

Several methods have been devised to generate random genetic codes for analyzing code properties (18–20), and different generators have yielded different results in some cases (19, 21, 22). To determine if our findings were independent of the code-generation methods, we employed all three methods in our correlation analyses. The first method (RAND method) maintains the number of codons per amino acid as dictated by the canonical code and randomizes the positions of the amino acids (19) (*Methods*). The second method (NNY method) is similar to the RAND method but imposes the restriction of not splitting NNY blocks (21). No evidence exists in nonstandard genetic codes that suggest translation can distinguish between codons occupying a NNY block (23), and thus it has been proposed that NNY codons should not be split when generating random codes (21). The third method (SYN method) maintains the exact block structure of the canonical code and randomizes the amino acid in each block (20). We generated one million codes using each method to perform the correlation analyses, and found that the results are consistent regardless of which method was employed (Table S1). The RAND method produced more optimized codes and less significant probabilities. As such, the RAND generator was used for all analyses described, as it was the most statistically stringent method.

Evidence for Codon Reassignment During the Evolution of the Canonical Code. In the canonical code, eight blocks of four codons that differ only at the third position are each shared by two amino acids or by amino acids with stop signals; the other eight blocks are each occupied by a single amino acid. Codon block sharing has implications on codon reassignment (23). To explore whether ribosomal anticodon-amino acid enrichment can reveal evidence for such potential evolutionary events, we used global correlation analysis to reanalyze the code by expanding each amino acid in the split blocks to occupy the entire codon block. The average enrichment value of each expansion of the canonical code was calculated and compared with that of each random code after expansion of the same amino acid. The number of "better" codes, which have a higher average enrichment value and are thus more correlated to the ribosome structure data than the canonical code, was determined before and after expansion of a particular amino acid.

Upon expansion to occupy a full codon block, Leu, Ile, His, and Lys all markedly improved the correlation of the expanded code with the ribosomal anticodon-amino acid enrichment, and thus had a large decrease in the percentage of better random codes (Fig. 3). In contrast, the other amino acid in each block (Phe, Met, Gln, and Asn, respectively) had no such large decrease after expansion. Remarkably, expansion of Phe showed the opposite effect, greatly increasing the number of better random codes by 40%. The other four shared-codon blocks did not show such contrast in change of better random codes (Fig. S1). These results suggest that Leu, Ile, His, and Lys may have once occupied the entire codon block, and specific codons were

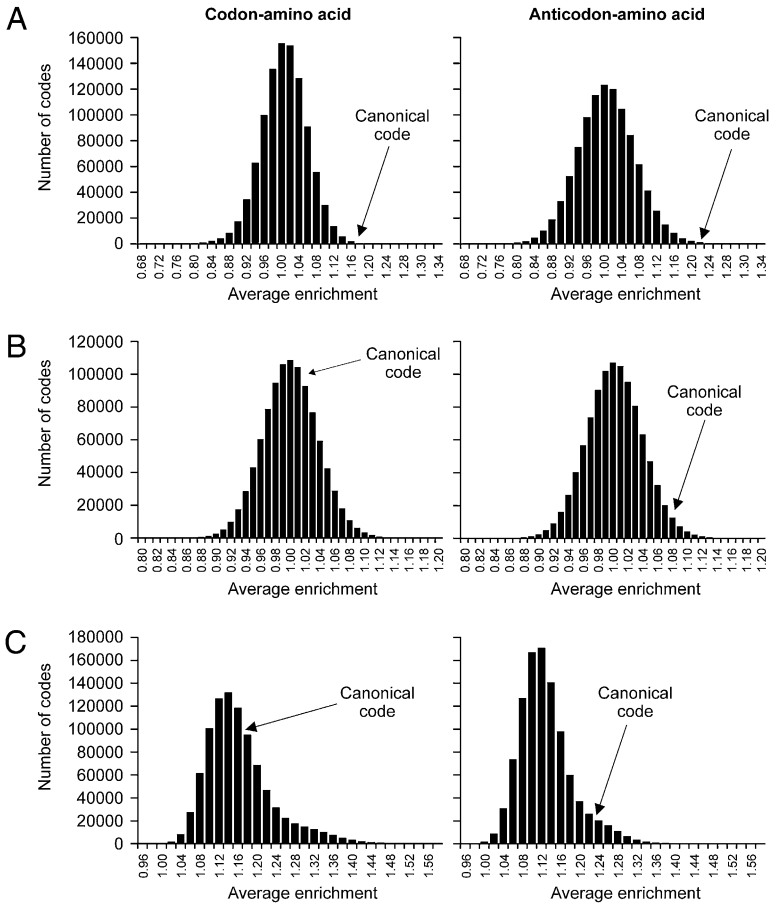


Fig. 2. Anticodon-amino acid enrichment observed in ribosomal structures is correlated with the canonical genetic code. Monte Carlo simulations were used to analyze the canonical code and random codes for codon- or anticodon-amino acid enrichment in the ribosomal structures. The distributions of the average enrichment value of codon (*Left*) or anticodon (*Right*) for 10^6 random codes are shown with the arrow pointing to where the canonical code stands. The correlation analyses were performed using (*A*) a specific subset of amino acids that is optimal for the canonical code as in Fig. 1 *D* and *E*; (*B*) all amino acids except cysteine; and (*C*) a subset of amino acids optimal for each code. A significant proportion of randomized codes have a lower average enrichment than the canonical genetic code in ribosomal anticodon-amino acid enrichment.

later conceded to the other amino acids. Specifically, the data provide evidence for the capture of AUG from Ile to Met, consistent with what is considered a key codon capture event (24, 25), as well as unique evidence that Phe may have captured the UUY codons from Leu. Our data also suggest that a subset of codons of Lys and His were reassigned to Asn and Gln, respectively.

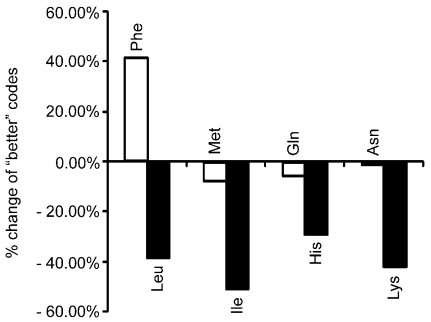


Fig. 3. Ribosomal anticodon-amino acid enrichment reveals evidence for codon reassignment during code evolution. Four amino acids in the shared codon blocks (Leu, Ile, His, and Lys), when each expanded to occupy the entire codon block, markedly improved the correlation of the code with the ribosomal anticodon-amino acid enrichment. The number of "better" codes, random codes with higher global average enrichment, was significantly reduced. The other amino acid in the shared codon blocks showed no such reduction when expanded.

Although the correlation between the canonical code and the ribosomal anticodon-amino acid enrichment is strong, some random codes were found to have a higher average enrichment value in the global correlation analysis. Analysis of amino acid placements in the top 1,000 random codes correlated with the ribosomal enrichment data may also reveal further evidence for reassignments of amino acids in code evolution. For example, Ile shows a strong preference for the Met codon AUG in 44% of the top 1,000 random codes (Fig. S2*A*). Consistently, CAUC, which contains the corresponding anticodon CAU, was found significantly enriched within 5 Å of Ile than the other 19 amino acids ($P = 0.0003$) in the ribosome. This result further suggests that Ile may have initially occupied the AUG codon and Met captured it later. In another example, 26% of the top 1,000 random codes place Pro at the AAN position (Fig. S2*B*) because of the significant enrichment of the UU moiety within 5 Å of Pro in the ribosome ($P = 0.0091$). Proline thus may have undergone a shift from AAN to its current CCN positions in the canonical code, for which no previous evidence exists.

Evidence of codon reassignment events have been previously limited to changes in nuclear and mitochondrial codes of certain organisms (23, 26). The ribosomal data shown here present unique evidence for the existence of these evolutionary events that drive the establishment of the canonical code. The reassignment of amino

acids to different codons during evolution also helps to explain why only a subset of amino acids are involved in amino acid-anticodon enrichment in the ribosome.

The Canonical Code Evolved Through a Two-Stage Transition. Our finding on ribosomal anticodon-amino acid enrichment, when compared with the consensus chronology of amino acids built on 60 criteria (24), revealed a distinct stage for the evolution of the canonical code. Except for Asp, all other amino acids that are significantly enriched with their anticodons in the ribosome are at the later stage of amino acid expansion (Table 1), starting from Leu (ranked No. 9) to Trp (No. 20). For the two amino acids at the later stage without apparent anticodon enrichment, Cys is too rare on the ribosomal protein surface for statistical analysis, leaving Asn as the only inconclusive amino acid. Lysine-rRNA interactions are the second most abundant amino acid-RNA interaction in the ribosome because of abundant ionic interactions with the RNA backbone, which masks the enrichment of Lys to its anticodons. However, with an apparent weak enrichment, Lys showed a marked decrease in the number of better random codes when expanded to occupy the full AAN block (Fig. 3). We therefore tentatively place lysine in the population of those amino acids with an anticodon interaction. The two-stage distribution suggests that ancient amino acids implied in the Miller-Urey experiment (27) were fixed into the genetic code through means other than amino acid-anticodon interaction, or that such interactions were masked in the ribosome because of later codon reassignment upon addition of new amino acids. Newer amino acids, on the other hand, were added to the genetic code mainly through the anticodon interaction.

Discussion

The canonical genetic code is one of the most dominant aspects of life on this planet, and thus studying the origin of the genetic code is critical to understanding the evolution of all life. In this study, we tried to address this long-standing and difficult question using structural information of ribosomes from bacteria and archaea, two of the three life domains. We found that anticodons are selectively enriched near their respective amino acids in ribosome structures and that such enrichment is correlated with the canonical genetic code. Previously, codons and anticodons have been found enriched in the binding sites in RNA aptamers selected to bind amino acids (4, 28), and conserved arginine codons have been identified to bind Arg specifically in group I self-splicing introns (11). Such experiments have provided precious, hard to obtain empirical data, demonstrating the association of coding triplets with amino acids. Our findings corroborate such association but place the anticodon-amino acid interactions in a meaningful biological context: modern ribosomes, the machinery responsible for translating the genetic code. Although selective pressures have continued refining the ribosome during evolution, our results suggest that the essence of the primitive RNA-amino acid interaction remains at the heart of modern ribosomes (29). More importantly, our data establish a direct connection between the anticodon-amino acid association and the genetic code, filling a critical gap that has not been addressed using any means before. Our findings are thus unique, providing comprehensive in vivo evidence supporting the tenets of the stereochemical hypothesis for code origin.

Table 1. Ribosomal anticodon-amino acid enrichment suggests a two-stage evolution of the canonical code

Order	Amino acid	Miller–Urey	Anticodon
1	G	+	
2	A	+	
3	D	+	+
4	V	+	
5	P	+	
6	S	+	
7	E	+	
8	T	+	
9	L	+	+
10	R		+
11	N		
12	I	+	+
13	Q		+
14	H		+
15	K		+
16	C		N/D
17	F		+
18	Y		+
19	M		+
20	W		+

The chronological order of amino acid addition to the code is shown, as well as the amino acids that are thought to have been prebiotically available in the Miller–Urey experiments (27). The amino acids with significant enrichment of anticodon-containing rRNA in ribosome structures are later additions to the genetic code.

The original stereochemical hypothesis proposes that the genetic code is shaped by interactions between amino acids and cognate coding triplets, which can be codons or anticodons. Whether codons or anticodons played the role in the origin of the genetic code has remained elusive (30, 31). In the in vitro experiments of selecting RNA aptamers to bind amino acids, both codons and anticodons are found significantly enriched in the cognate amino acid binding sites. Specifically, Arg and Ile are found to associate with both codon and anticodons, whereas His, Phe, Trp, and Tyr are associated with anticodons only (28). Interpreting the stereochemical complementarity for amino acids enriched with both codons and anticodons in RNA aptamers is difficult. In the arginine binding site of the self-splicing group I introns, only codons are found to interact with Arg (11). In contrast, our results clearly indicate that it is the anticodon-amino acid association, but not the codon-amino acid association, correlated with the genetic code. This conclusion advances the stereochemical hypothesis by pinpointing anticodon-amino acid interaction as the potential drive for code formation. In addition, this finding will help our understanding of the origin of the translation system. It has been proposed that one critical step is the evolving of amino acid-binding RNAs under the selection pressure of amino acid accumulation (31). The specific binding mechanism determines the correspondence between amino acids and cognate triplets. These RNAs subsequently acquire autocatalytic aminoacylation to covalently couple the amino acid and eventually become tRNAs. The anticodon-amino acid binding fits naturally with the evolution and function of the tRNA, which contains the anticodon. Further study based on the anticodon-amino acid association will shed new light onto the predecessors and elementary steps that eventually led to the emergence of translation.

There are two distinct periods for the evolution of the genetic code, before and after the LUCA. The canonical code emerged before the LUCA; all life following the LUCA inherited this canonical code nearly ubiquitously, yet minor deviations from the canonical code have been found in some nuclear and mitochondria lineages (26). Such code variants suggest that the genetic code is still evolving and codon meanings can be reassigned in extant organisms. Had codons also been reassigned during the establishment of the canonical code before LUCA? Both the adaptation theory and the coevolution theory hypothesize codon reassignment as a necessary step for the formation of the canonical code. The adaptation theory predicts that the code underwent optimization to minimize the impact of errors of translation (32); the coevolution theory predicts that the code was expanded through the addition of biosynthetically related new amino acids, which took codons from the precursor amino acids (33). However, because primitive life before LUCA and LUCA itself do not exist today for comparative study, evidence for codon reassignments during the canonical code formation is lacking. Our results, obtained through analyzing ribosomal structural

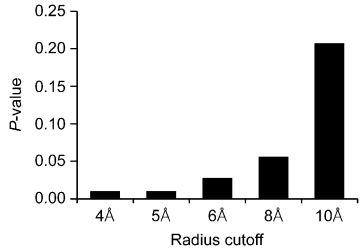


Fig. 4. Determination of the radius cutoff used for all analyses. Amino acid-rRNA interactions in ribosome structures were defined using radii ranging from 4 to 10 Å. Global correlation analysis of ribosomal anticodon-amino acid enrichment and the code was performed on 10^6 random codes using these five datasets to determine the *P* values corresponding to each radius. Similar analysis of ribosomal codon-amino acid enrichment and the code was not statistically significant.

data and its correlation with the canonical code, now provide such evidence. Specifically, we found evidence suggesting that a subset of codons originally set for Leu, Ile, His, and Lys were reassigned to Phe, Met, Gln, and Asn, respectively, and that Pro was reassigned from AAN to the current CCN codons. Compared with the codon reassignments found in post-LUCA organisms (23), the pre-LUCA codon reassignments we identified show overlaps and distinctions: The reassignment events from Ile to Met and from Lys to Asn have also been observed in mitochondria and nuclear lineages, whereas those from Leu to Phe, from His to Gln, and for Pro have never been identified in extant life. It should be noted that the codon reassignments identified in this study may not cover all possible reassignment events. We could not analyze the reassignment between stop codons and sense codons because of the nonexistence of a stop entity in proteins, whereas the reassignment of stop codons to amino acids is often seen in mitochondrial and nuclear code variants.

It is hypothesized that the coding process started with a set of primitive amino acids and that others were added until the total of 20 was reached. What amino acids were the primordial ones and how the new amino acids were added to the genetic repertoire are unknown and under debate (4, 19, 21, 34). In addition, if RNA-amino acid interactions influenced the organization of the code, the pervasiveness of these interactions, and the interplay between such interactions and other potential driving forces, are also unknown. A unique finding from our study is the temporal order in which anticodon-amino acid interactions may have taken place. We demonstrate that the canonical code may have evolved through two distinct stages. The formation of the early code, consisting of the prebiotically available amino acids, was not influenced by the anticodon-amino acid interactions. Once some critical events occurred—for example, the establishment of primitive translation—these interactions became relevant and induced a watershed for code expansion, allowing newly available amino acids a route into the genetic code.

A major challenge in studying the origin of the genetic code is that many questions are out of reach of direct experimentation. Empirical data, although limited in the past, has yielded landmark insights. The spark-tube experiments confirmed that amino acids could have been produced abiotically in the atmosphere of the early earth (27, 35), and in vitro RNA aptamer selections provided solid evidence for the association of coding triplets with cognate amino acids (28). Our usage of ribosome structural data further demonstrates the usefulness of expanding the theoretical inputs to a biological context. This methodology leads to robust results and unique findings, underscoring the potential created by empirical evidence in the study of code origin and evolution.

Methods

Ribosome Structural Analysis. Ribosome structural data were analyzed by using new Perl scripts. If any atom of an amino acid was within 5 Å of an rRNA nucleotide, this was considered a hit and the amino acid and surrounding rRNA sequence was put into a separate array. The arrays from all four structures were compiled and used for all subsequent analyses.

The selection of 5 Å as the cutoff was the result of balancing two conflicting parameters: number of interactions and interaction specificity. RNA-amino acid interactions would be more specific within a closer distance, and thus a smaller radius would allow more accurate analysis. However, the number of interactions would become limiting if the radius is too small, leading to the exclusion of certain amino acids from robust statistical analyses. To determine the cutoff, global correlation analysis (see below) was performed on data generated from radii ranging from 4 to 10 Å. The lower limit was set by the resolution of the ribosome structures themselves. As expected, an increased radius led to less significant results, with only 4 to 6 Å radii having a P value $<$ 0.05 (Fig. 4). Radii of 4 or 5 Å had very similar statistics; however, the limited data from 4 Å led to the exclusion of three amino acids. Data generated from a 5 Å radius only excluded cysteine (see below), and was therefore used for all subsequent analyses.

Probability Analysis of Anticodons or Codons. Populations containing every possible three-nucleotide RNA sequence were extracted from the data using new Perl scripts and used to calculate probabilities. The probability for finding a certain sequence near a specific amino acid was then calculated for all anticodon- and codon-containing sequences for all 20 amino acids. The probability (P_X where X is the amino acid in question) of finding a particular set of codon- or anticodon-containing sequences within 5 Å of an amino acid versus all other sequences was calculated. For those amino acids with multiple codons or anticodons, the results were additive. For example, Phe in the canonical code has two codons, UUU and UUC. P_F would be the number of times UUU and UUC appears within 5 Å of Phe residues divided by the number of total three-nucleotide sequences (NNN) within 5 Å of Phe. A normalization factor (N_X where X is the amino acid in question) was then calculated using the exact same calculations, only for the 19 remaining amino acids. The calculation for N_F in this case would be the number of times UUU and UUC appear within 5 Å of the other 19 amino acids divided by the number of total three-nucleotide sequences (NNN) within 5 Å of the 19 amino acids. The final enrichment value for Phe would then be $E_F = \frac{P_F}{N_F}$. This enrichment value thus represents the increase or decrease in likelihood of finding a particular set of three-nucleotide sequences near a particular amino acid. If the value is 1, then there is no increased likelihood of finding these sequences near a particular amino acid, compared with the other 19 amino acids.

A two-tailed χ^2 analysis was then used to determine if any codon- or anticodon-containing sequences were more likely to be found near amino acid X versus the other 19 amino acids. If there were fewer than 10 hits for any given sequence, the statistical analysis is less robust and was thus not determined. All amino acids showing enrichment were then sorted by their P value for combining results via Fisher's method. The combination of the largest set of amino acids that satisfied $P < 0.05$ was determined and used in the subsequent Monte Carlo simulations.

Generation of Random Codes. Each of the three methods for generating 1,000,000 randomized codes used a series of new Perl scripts. The RAND generator was essentially as described by Novozhilov et al. (19), with a few minor modifications. In the RAND and NNY generators we did not maintain the positioning of the stop codons, and each was assigned independently. These generators put more emphasis on maintaining the number of codons per amino acid, and less on maintaining the exact structure of the canonical code. In addition, Ile and Met were not treated as a single block and were assigned independently. Those amino acids occupying a full block (four codons) were first assigned a position in the 16 possible locations using Perl's random number generator. Ile was then assigned by first treating it as occupying two codons and was placed randomly in any of the remaining NNY or NNR blocks. Following placement, the third Ile codon was randomly placed in the corresponding neighboring block (i.e., if Ile was assigned to AAY, then the third position would be AAA or AAG). Next, the amino acids with two codons were randomly assigned to any remaining NNY or NNR blocks, followed by those that occupy a single codon and the three stop codons. Those amino acids with six codons were split up into a four-codon block and a two-codon block, which were assigned independently.

The NNY generator was designed after the RAND generator, but with modifications to ensure that no NNY block was split. Ile was first randomly paired with Met, Trp, or a stop codon, and then the entire block was randomly assigned codons using Perl's random-number generator. There was no restriction on which NNR the third Ile codon occupied. The four remaining single codon-containing amino acids and stop codons were randomly paired

and assigned to a random NNR. The remaining four-codon and two-codon amino acids were then assigned as described in the RAND generator.

The SYN generator was designed as in Haig and Hurst (20). With this generator, we did not alter the position of the stop codons, as the impetus behind the creation of these random codes is different from those above. This generator assumes that the exact structure of the code is critical to its function, and in theory, any major deviation would make the randomized codes more unrealistic and unfit. To create these codes, each block as it exists in the canonical code was assigned a number. Each of the 20 amino acids was then randomly assigned to a block to ensure that each amino acid was represented at least once. The remaining three blocks were then randomly assigned any of the 20 amino acids, completing the randomized codes.

Comparison of Random Codes. For all comparisons, average enrichment was chosen over χ^2 analysis because it is less likely to be influenced by a single member of the population. Average enrichment was also much faster to calculate and thus facilitated the running of the numerous simulations required. Finally, results between the two analyses should be very similar regardless, as the two numbers themselves are significantly correlated ($P < 0.0001$) using the nonparametric Spearman calculation.

For the specific subset analysis, only those amino acids determined from the canonical code using the Fisher's method were analyzed for each randomly generated code. The enrichment values for the codons or anticodons, as dictated by the randomized code, were calculated for each of the amino acids. The average enrichment was calculated and compared with that of the canonical code. Those codes with a higher average enrichment were considered more correlated with the ribosome structural data.

For both the global and optimal subset analysis, all amino acids were analyzed from each randomly generated code. The enrichment of either codons or anticodons as dictated by the randomized code were calculated and sorted. For the global analysis, an average enrichment value for the 22 independent calculations was obtained and compared with the canonical code. Those amino acids with six codons were split into their respective four- and two-codon blocks and analyzed independently. Cys was removed because it contained the fewest number of surrounding rRNA hits, leading to an increase in the percent error of the randomized codes (Fig. S3). In addition, Cys rarely met the >10 hits cutoff imposed on the statistical analysis in the majority of the random codes. For the optimal subset, those amino acids that had the highest 8 enrichment values for codons or 11 for anticodons were used for each randomized code. Therefore, each randomized code may have a different subset of amino acids used for comparison. Again, those codes with a higher average global or optimal enrichment are considered more correlated with the ribosome structural data.

Analysis of the Shared Codon Blocks. Each amino acid in a shared block was expanded individually and the global Monte Carlo analysis was performed and compared with the standard global anticodon analysis before expansion. Any difference in the number of more correlated random codes was used as a comparison to determine if expansion led to a higher correlation. In the analysis of the shared codon blocks, one amino acid was expanded to occupy the full codon block; the other amino acid sharing the block could be either analyzed at its positions in each code or excluded from the analysis. Exclusion of the sharing amino acid makes direct comparison of random codes difficult because the identity of the sharing amino acid may vary in different codes. For example, in the case of the canonical code, the expansion of Ile would measure the enrichment of all NAU anticodons for Ile and exclude the calculations for Met. In random codes using the RAND generator, Ile is generally paired with Met, Trp, or a stop codon. In these codes, the excluded amino acid could be any of those possibilities. Exclusion of different amino acids makes comparison of two codes ambiguous and potentially incorrect, as the makeup of the calculations would be inconsistent. For these reasons, we calculated the sharing amino acid at its positions in each code. This method led to one or two anticodons being used twice for each calculation. For example, in the Ile expansion analysis of the canonical code, Ile occupied the full NAU block and Met occupied CAU. This method allowed us to never exclude an amino acid and to make more accurate comparison of random codes.

ACKNOWLEDGMENTS. We thank Drs. Joseph Noel and Tony Hunter for helpful discussions. This work was supported by the Searle Scholar Program (06-L-119), the Beckman Young Investigator Program, and the National Institutes of Health Director's New Innovator Award DP2OD004744 (to L.W.).

1. Szathmáry E, Smith JM (1995) The major evolutionary transitions. *Nature* 374: 227–232.
2. Woese CR (1965) Order in the genetic code. *Proc Natl Acad Sci USA* 54:71–75.
3. Knight RD, Freeland SJ, Landweber LF (1999) Selection, history and chemistry: The three faces of the genetic code. *Trends Biochem Sci* 24:241–247.
4. Yarus M, Caporaso JG, Knight R (2005) Origins of the genetic code: The escaped triplet theory. *Annu Rev Biochem* 74:179–198.
5. Woese CR, Dugre DH, Saxinger WC, Dugre SA (1966) The molecular basis for the genetic code. *Proc Natl Acad Sci USA* 55:966–974.
6. Majerfeld I, Yarus M (1998) Isoleucine:RNA sites with associated coding sequences. *RNA* 4:471–478.
7. Mannironi C, Scerch C, Fruscoloni P, Tocchini-Valentini GP (2000) Molecular recognition of amino acids by RNA aptamers: The evolution into an L-tyrosine binder of a dopamine-binding RNA motif. *RNA* 6:520–527.
8. Legiewicz M, Yarus M (2005) A more complex isoleucine aptamer with a cognate triplet. *J Biol Chem* 280:19815–19822.
9. Dunnill P (1966) Triplet nucleotide-amino-acid pairing; A stereochemical basis for the division between protein and non-protein amino-acids. *Nature* 210:1265–1267.
10. Majerfeld I, Yarus M (2005) A diminutive and specific RNA binding site for L-tryptophan. *Nucleic Acids Res* 33:5482–5493.
11. Yarus M, Christian EL (1989) Genetic code origins. *Nature* 342:349–350.
12. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* 289:905–920.
13. Fox GE, Naik AK (2004) *The Genetic Code and the Origin of Life*, ed de Pouplana LR (Landes Bioscience, New York), pp 92–105.
14. Selmer M, et al. (2006) Structure of the 70S ribosome complexed with mRNA and tRNA. *Science* 313:1935–1942.
15. Borovinskaya MA, et al. (2007) Structural basis for aminoglycoside inhibition of bacterial ribosome recycling. *Nat Struct Mol Biol* 14:727–732.
16. Wilson DN, et al. (2008) The oxazolidinone antibiotics perturb the ribosomal peptidyl-transferase center and effect tRNA positioning. *Proc Natl Acad Sci USA* 105:13339–13344.
17. Schmeing TM, Huang KS, Kitchen DE, Strobel SA, Steitz TA (2005) Structural insights into the roles of water and the 2′ hydroxyl of the P site tRNA in the peptidyl transferase reaction. *Mol Cell* 20:437–448.
18. Freeland SJ, Hurst LD (1998) The genetic code is one in a million. *J Mol Evol* 47: 238–248.
19. Novozhilov AS, Wolf YI, Koonin EV (2007) Evolution of the genetic code: Partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *Biol Direct* 2:24.
20. Haig D, Hurst LD (1991) A quantitative measure of error minimization in the genetic code. *J Mol Evol* 33:412–417.
21. Ronneberg TA, Landweber LF, Freeland SJ (2000) Testing a biosynthetic theory of the genetic code: Fact or artifact? *Proc Natl Acad Sci USA* 97:13690–13695.
22. Amirnovin R (1997) An analysis of the metabolic theory of the origin of the genetic code. *J Mol Evol* 44:473–476.
23. Knight RD, Freeland SJ, Landweber LF (2001) Rewiring the keyboard: Evolvability of the genetic code. *Nat Rev Genet* 2:49–58.
24. Trifonov EN (2004) The triplet code from first principles. *J Biomol Struct Dyn* 22:1–11.
25. Jukes TH (1973) Possibilities for the evolution of the genetic code from a preceding form. *Nature* 246:22–26.
26. Osawa S, Jukes TH, Watanabe K, Muto A (1992) Recent evidence for evolution of the genetic code. *Microbiol Rev* 56:229–264.
27. Miller SL, Urey HC, Oró J (1976) Origin of organic compounds on the primitive earth and in meteorites. *J Mol Evol* 9:59–72.
28. Yarus M, Widmann JJ, Knight R (2009) RNA-amino acid binding: A stereochemical era for the genetic code. *J Mol Evol* 69:406–429.
29. Woese CR (2001) Translation: In retrospect and prospect. *RNA* 7:1055–1067.
30. Szathmáry E (1999) The origin of the genetic code: Amino acids as cofactors in an RNA world. *Trends Genet* 15:223–229.
31. Wolf YI, Koonin EV (2007) On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization. *Biol Direct* 2:14.
32. Freeland SJ, Wu T, Keulmann N (2003) The case for an error minimizing standard genetic code. *Orig Life Evol Biosph* 33:457–477.
33. Wong JT (1975) A co-evolution theory of the genetic code. *Proc Natl Acad Sci USA* 72: 1909–1912.
34. Di Giulio M, Amato U (2009) The close relationship between the biosynthetic families of amino acids and the organisation of the genetic code. *Gene* 435:9–12.
35. Miller SL (1987) Which organic compounds could have occurred on the prebiotic earth? *Cold Spring Harb Symp Quant Biol* 52:17–27.