AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119
(Some corrections may occur before final publication online and in print)

Annu. Rev. Biochem. 2005. 74:179–98
doi: 10.1146/annurev.biochem.74.082803.133119

First published online as a Review in Advance on February 18, 2005

ORIGINS OF THE GENETIC CODE: The Escaped Triplet Theory

Michael Yarus,[1] J. Gregory Caporaso,[2] and Rob Knight[2]
[1]Department of Molecular Cellular and Developmental Biology, [2]Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado 80309-0347; email: yarus@stripe.colorado.edu, caporaso@colorado.edu, rob.knight@colorado.edu

Key Words RNA, selection, Bayes' theorem, translation, evolution

■ **Abstract** There is very significant evidence that cognate codons and/or anticodons are unexpectedly frequent in RNA-binding sites for seven of eight biological amino acids that have been tested. This suggests that a substantial fraction of the genetic code has a stereochemical basis, the triplets having escaped from their original function in amino acid–binding sites to become modern codons and anticodons. We explicitly show that this stereochemical basis is consistent with subsequent optimization of the code to minimize the effect of coding mistakes on protein structure. These data also strengthen the argument for invention of the genetic code in an RNA world and for the RNA world itself.

CONTENTS

11 Feb 2005 16:55 AR AR261-BI74-07.tex XMLPublish^{SM}(2004/02/24) P1: JRX
AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119

INTRODUCTION

The genetic code, conserved with small variations (1) in the 10s of 1000s of organisms for which sequence data are available, is one of the hallmarks of life on Earth. Assuming only that codons could have been assigned differently (discussed below), this uniformity becomes telling evidence that the competition between early living forms was won by a single group of creatures (2) whose descendants are the only clade of life known on Earth (indeed, anywhere in the universe). It is of great interest to understand the emergence of this near-ubiquitous property of life. Three ideas about the origins of the present system of genetic coding are supported by substantial empirical arguments, and we contrast these below.

Stereochemical Origin

We will argue that this hypothesis (3) is supported by the most extensive evidence. It states that there is specific affinity between codons or anticodons and amino acids. The stereochemical hypothesis is an idea with an intrinsic experimental program because it predicts that the genetic code relies on interactions that should be observable in a modern laboratory. We have tested the idea that the coding triplets arise as essential parts of RNA-like amino acid–binding sites. These ancestral triplets then escape from amino acid–binding sites to acquire new functions as codons and anticodons in a more modern translation apparatus. It is not required that all associations in the coding table have this character (4), and in particular, we know that the code can change (see below). The original coding assignments might thus have been overwritten, using other logic, during later code evolution. Nevertheless, it seems that many chemical attractions, manifested as binding affinities, still seem to be reflected in the modern coding table.

Adaptation

This is the idea that the genetic code has been refined to minimize the impact of coding errors (5) or the likelihood that mutated proteins would be active (3). The recent discussion of this question was initiated by the finding of Freeland & Hurst (6) that, weighted to account for translational error, the real code is better than all but one of a million random codes in minimizing chemical differences (hydrophobicity) between amino acids whose codons are linked by single-nucleotide misreadings. There is a recent review of the entire topic by Freeland (7). Because adaptation requires a starting code to be refined, it explicitly suggests a more ancient nucleus of coding assignments made on another basis. In fact, if the initial nucleus of the code were stereochemical, organisms that escaped from the chemical constraints imposed by stereochemical assignment (for instance, by using tRNA to decouple codon recognition from affinity for the amino acid) might have more scope for the free reassignment required for adaptation, and these organisms might be more likely to outcompete their less flexible relatives. Thus adaptation of the code would be expected after the time of the most ancient encoding. The available

11 Feb 2005 16:55 AR AR261-BI74-07.tex XMLPublish^{SM}(2004/02/24) P1: JRX
AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119

evidence suggests that a substantial fraction of the modern code is composed as if it were fixed by adaptive selection on amino acid hydrophobicity, possibly to conserve protein folding. However, some of the apparently adaptive assignments may also be explicable by the stereochemical or coevolution hypotheses: All three predict that similar amino acids will be assigned to similar codons. The process of adding new amino acids to the genetic code in an error-minimizing fashion could, according to modeling, have given the code its overall shape [for example, with chemically related amino acids in coding table columns (7, 8)].

Coevolution

In the description of Wong (9), an initial code, containing a nucleus of codons and primordial, nonbiosynthetic amino acids, is extended to amino acids whose biosynthesis later becomes possible. New metabolic products obtained possession of some of the codons previously allotted to their molecular precursors in the broadly used anabolic pathways. This is again an idea that necessarily specifies a heterogeneous origin for the code, and it requires a nucleus of assignments that are elaborated as metabolism expands. There has been a lively recent discussion of coevolution initiated by the work of Amirnovin (10), who pointed out that the choice of precursor-product pairs was crucial to the apparent plausibility of coevolutionary shaping of the coding table. For example, if pairs from *Escherichia coli* instead of supposed universal pairs were used, the ability to predict amino acids related by single-base changes mostly disappeared. This conclusion can itself be criticized owing to biased selection of related amino acid pairs (11). However, Ronneberg et al. reached the same conclusion (12) on the basis of a reanalysis of usable precursor-product amino acid pairs using pathway topology and thermodynamic interrelations between the compounds. The same authors pointed out that codons NNU and NNC are rarely distinguished in translation: If they are instead regarded as synonymous, the apparently adjacent location of the codons of precursor-product amino acid pairs loses significance. The coevolution hypothesis has been redefended by Di Giulio (13), for example, by giving codons to plausible intermediate amino acids homoserine, ornithine and citrulline, but supposing that they are no longer encoded. This adds hypotheses for which there is no direct evidence, and it seems to us that the present code can no longer be said to statistically support coevolution—the choice of the valid primordial precursor-product amino acid pairs is both influential and uncertain.

However, it is important to be clear about what is at stake in this discussion. This negative conclusion, even if accepted completely, undermines only the idea that the majority of the code coevolved using the specific pathwaysoriginally suggested by Wong (9). The evolutionary concession of certain codons to later-appearing, synthetically related amino acids remains an appealing idea. In particular, as has been pointed out by Di Giulio (11), the observed conversion of Glu-tRNA to Gln-tRNA (14) and the conversion of Asp-tRNA to Asn-tRNA (15) can be interpreted as striking molecular remnants of the codon concession process exactly as originally envisioned in the original coevolution theory of Wong (9).

11 Feb 2005 16:55 AR AR261-BI74-07.tex XMLPublish^{SM}(2004/02/24) P1: JRX
AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119

The Three Hypotheses Could All Be Correct

Adaptation and coevolution are, in generalized form, both potentially compatible with stereochemical origins for some codons because each requires a starting code from sources potentially distinct from its own mechanism. The three processes are complementary rather than competitive (4) and could have occurred simultaneously, even though they are often posed as alternatives for polemical purposes.

THE CODE EVOLVES

It might be thought that the possibilities immediately above, in which the code is cast as a device subject to evolution, cannot be correct. Isn't the genetic code frozen (16) as the result of protein damage owing to a myriad of simultaneous amino acid changes? No, it is not; even modern organisms with 1000s of highly evolved proteins have changed the meaning of a few codons in their genetic codes (1). Change was probably even easier in more primitive creatures, with more error-prone translation, fewer gene products, and a lesser penalty for imprecision. Modern codon reassignments are not random but highly repetitive. It appears that this selective repetition strongly reflects the directions of underlying coding error (17), so we believe that ambiguous coding (codons with multiple meanings) provides the transitional state via which an original meaning for a codon shifts to a simultaneous new one. Modern organisms can switch codes via ambiguous transitional periods, either following drift in genome composition [and the resulting codon bias and low codon frequencies (1)] or in a manner independent of drift in composition (18). The code, therefore, is not frozen but perhaps chilled; selection against new amino acids in many sites slows down evolutionary change so new codes stay in a selective basin near the universal one. However, there is no logical or empirical difficulty in positing a history for the genetic code in which many variations were tried [reviewed in (19)].

The Current State of the Stereochemical Argument

We now summarize the experimental data that lead us to a stereochemical conclusion: Triplet codon and anticodon sequences from amino acid–binding structures made of something very like RNA were used to build the genetic code. Because so much has been written in the past on this topic, we can readily leave some parts of the argument to the literature. Anyone interested in the origins of the code should start with the classics by Crick (16) and Orgel (20). The RNA world and the ribocyte are reviewed by Joyce (21) or at more varied length in the eponymous book (22). Recent reviews summarize the evidence for the origin of the entire translation apparatus, including the code, in RNA chemistry (23) and broader implications of RNA chemistry for the RNA world hypothesis (24). Most specifically, the experimental argument for the escaped triplet hypothesis was assembled from data available at an earlier stage (25). The fundamental observation is that RNA molecules selected from random sequences to bind specific amino acids have far

TABLE 1 Probabilities for experimental associations between codons and binding sites, anticodons and binding sites, and both codons and anticodons with binding sites for each amino acid separately, and all amino acids together (27)

Amino acid	Codon (C)	Anticodon (AC)	Both C & AC	Aptamers/Nucleotides
Phenylalanine	0.72[a]	3.4×10^{-5}	2.9×10^{-4}	2/160
Isoleucine	1.2×10^{-3}	1.0×10^{-6}	2.7×10^{-8}	6/320
Histidine	0.999	6.9×10^{-4}	5.7×10^{-3}	12/809
Leucine	0.27	4.5×10^{-4}	1.2×10^{-3}	1/78
Glutamine	0.042	0.99	0.17	2/156
Arginine	3.4×10^{-8}	0.045	3.3×10^{-8}	5/197
Tryptophan	0.99	1.8×10^{-4}	1.7×10^{-3}	12/800
Tyrosine	4.8×10^{-3}	1.6×10^{-6}	1.6×10^{-7}	3/271
Overall	6.6×10^{-3}	1.1×10^{-10}	5.4×10^{-11}	43/2791

[a]Basic calculations are performed as in (62); Fisher's method for combining multiple hypothesis tests is used to combine probabilities across amino acids and between codons and anticodons. The significance of individual observations is assessed by requiring that the probability of erroneous association between triplets and sites be $\leq$0.01 for the entire discussion, with no individual errors of assignment. But to insure that there will be less than 1 error in a 100 after 8 comparisons are done, errors for individual amino acids must have a probability p in $1-(1-p)^8 \leq 0.01$. This implies that probabilities of 1.3×10^{-3} and smaller are significant for individual amino acids (with 10^{-2} and smaller for overall probabilities). References for the aptamers are phenylalanine, (63; M. Illangasekare and M. Yarus, in preparation); isoleucine, (29; M. Legewicz and M. Yarus, in preparation); histidine, I. Majerfeld, D. Puthenvedu, and M. Yarus, in preparation; leucine, I. Majerfeld and M. Yarus, in preparation; glutamine, C. Scerch and G.P. Tocchini-Valentini, personal communication; arginine, (62, 64–71); tryptophan, I. Majerfeld and M. Yarus, in preparation; tyrosine, (72).

more of the coding sequences (codons or anticodons, or both) for those amino acids in the standard genetic code than would be expected by chance. By comparison with the last review (25), new data for isoleucine, histidine, phenylalanine, and tryptophan have been added. Overall, the probability of seeing a codon association as great as that observed is 6.6×10^{-3}, and the comparable figure for anticodons is 1.1×10^{-10} (Table 1; discussed below).

For this discussion, we revisit the occurrence of codons and anticodons in amino acid–binding sites isolated by selection amplification (SELEX), combining all data in our possession. We now know of 43 recently selected and characterized binding sites for 8 amino acids in RNAs with 2791 total initially randomized nucleotides. Our intent here is to use these data conservatively to arrive at minimal estimates for the probability of the calculated associations (26). To estimate the probability of observed associations between coding triplets and binding sites, we considered only those sequences for which direct experimental information about the binding sites was available (chemical protection/modification/interference mapping, conservation, or NMR). We considered only independently derived structures in which the binding site occurred in backgrounds with no significant sequence identity. For example, the minimal isoleucine site has been isolated at least 63 times, and the minimal histidine site has been isolated at least 54 times. However, only a minority

11 Feb 2005 16:55 AR AR261-BI74-07.tex XMLPublish^{SM}(2004/02/24) P1: JRX
AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119

of individually characterized examples are used for calculations in Table 1. Inclusion of other independent isolates would increase the statistical significance of the results (diminish the probabilities) by many orders of magnitude.

To calculate probabilities, we assigned each aptamer nucleotide to one of four categories: both triplet and binding-site nucleotide, triplet and not a binding-site nucleotide, not a triplet nucleotide but a binding-site nucleotide, or neither triplet nor binding-site nucleotide, depending on whether it was in a coding triplet (either a codon or an anticodon for the cognate amino acid) and whether it was in the experimentally determined binding site. These four counts were pooled across sequences for the same amino acid, and subjected to the G test for independence (with the Williams correction), which tests whether the proportion of binding site nucleotides that participate in coding triplets exceeds the proportion of nucleotides in flanking sequences that participate in coding triplets. The G test is similar to but more accurate for small samples than the chi-squared test traditionally used in genetics. Thus, the nonbinding parts of the aptamers act as an internal control for all manipulations and for the net nucleotide composition. Counts are pooled across codons or anticodons and across all aptamers, for one amino acidto increase sensitivity.

To compare results across multiple analyses (different amino acids, codons, and anticodons), we used Fisher's method for combining independent tests of a hypothesis. We combined the tests for codons and anticodons for each amino acid and combined the tests for all amino acids to get overall estimates of the probabilities bearing on the escaped triplet hypothesis (Table 1).

We emphasize a few aspects of these results. 1. Codons and cognate binding sites are not independent sequences, and the association between anticodons and newlyselected aptamer binding sites is even more unlikely. 2. Not all amino acids show codons or anticodons concentrated in their binding sites, and codons and anticodons are not necessarily associated. Arginine sites improbably contain only codons; tryptophan and histidine sites contain only an excess of anticodons. 3. For both codons and anticodons, there are significant arguments for more than one single amino acid. The argument for codons could rest on arginine or isoleucine alone, whereas the argument for anticodons could be based only on isoleucine, tyrosine, or tryptophan. Thus, the evidence for triplet association with binding sites does not rely completely on any particular selection, any particular amino acid, or indeed on experimentation by any one laboratory. 4. Conversely, the argument can be made from the most significant single amino acids without reference to pooled data—arginine for codons and isoleucine for anticodons would serve well. Thus, the major conclusions can be supported using homogeneous subsets of the data and do not emerge only by combining disparate experiments. 5. There are true negatives, such as glutamine anticodons and histidine codons, indicating that this set of techniques does not unexpectedly force a positive result. 6. Only one amino acid, glutamine, shows no significant associations with either codons or anticodons. Stereochemical associations seem the rule rather than the exception. 7. The amino acid sites showing strong associations with cognate triplets are

11 Feb 2005 16:55 AR AR261-BI74-07.tex XMLPublish^{SM}(2004/02/24) P1: JRX
AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119

chemically varied; basic, polar, aliphatic, and aromatic side chains are detected. 8. Positive triplets do not have any obvious sequence or compositional properties, being AU-rich, GC-rich, and mixed. 9. Controls using arbitrarily derived triplets with the same compositions, such as reversed codons (25), do not concentrate in amino acid–binding sites.

In summary, current evidence, construed in a way that minimizes its significance, supports significantly elevated codon concentration in the binding sites for isoleucine, and arginine as well as elevated anticodon concentration for six of the eight amino acids, with the exceptions of glutamine and arginine. It is intriguing that the one exception to stereochemical origin is glutamine, an amino acid for which the evidence of coevolution (see above) is among the strongest. The overall probability that codons and anticodons for eight amino acids are independent of their selected cognate binding sites (counting favorable and unfavorable cases together) is exceedingly low, 5.4×10^{-11}.

Although the numerical details may change with addition of data, the overall trends to association of codons with a few kinds of amino acids and anticodons with the sites for many or most amino acids appear secure by normal standards (Table 1). Because these sequences were selected requiring only affinity for free amino acids, there seems no simple or otherwise plausible alternative to the idea that progenitor structures that bound amino acids gave rise to the coding triplets of the present genetic code.

Another Statistical Approach

The above statistics, which reveal whether a particular coding triplet (or set of triplets) is surprisingly abundant in a cognate binding site, come from a straightforward procedure. If there is no association between triplets and binding sites, each triplet would be as abundant at positions outside the binding site as at positions within it. This test asks how surprising the observed aptamer sequences are, given the standard genetic code. One might instead ask how surprising the standard genetic code is, given the aptamer sequences. This latter question is addressed by constructing randomized genetic codes and measuring the fraction of genetic codes that have associations between coding triplets and binding sites at least as great as those observed in the actual genetic code. The attraction of this method is that it eliminates any effects resulting from peculiarities of the selected sequences, such as biases in composition and spatial correlations among binding site nucleotides. Using the smaller collection of amino acid–binding sites then available, we previously found that codons in the real code are associated more with real binding sites than in 99.2% to 99.96% of all randomized codes, in which the variation reflects the fact that some randomization methods preserved fragments of the actual code (25).

Inclusion of recently isolated amino acid sites changes these calculations because the selection of new amino acids has strengthened anticodon associations without adding to support for codon associations (Table 1). Using an expanded

11 Feb 2005 16:55 AR AR261-BI74-07.tex XMLPublish^{SM}(2004/02/24) P1: JRX
AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119

set of randomization models, we find that the real genetic code has greater codon associations than 83.6% to 98.6% of random codes (average 90.3%) and greater anticodon associations than 97.4% to 99.99% of random codes (average 99.8%) (27). Combining the evidence for codons and anticodons, the real code has greater associations with coding triplets than 99.6% to 99.99% of random codes (average 99.8%). These estimates are slightly less significant than previously because, as more of the code is covered by existing aptamers, there are more ways for randomized codes to contain relevant pieces of the actual code; this in turn makes the associations of the randomized codes appear better. Thus, this test is increasingly conservative as more aptamers are tested. Furthermore, randomization methods that allow fragments of the real code to recur account for the lower fractions. Nevertheless, the idea that the real genetic code is unusual with regard to cognate triplets in binding sites obtained from experiments is still strongly supported.

The Simplest Sites

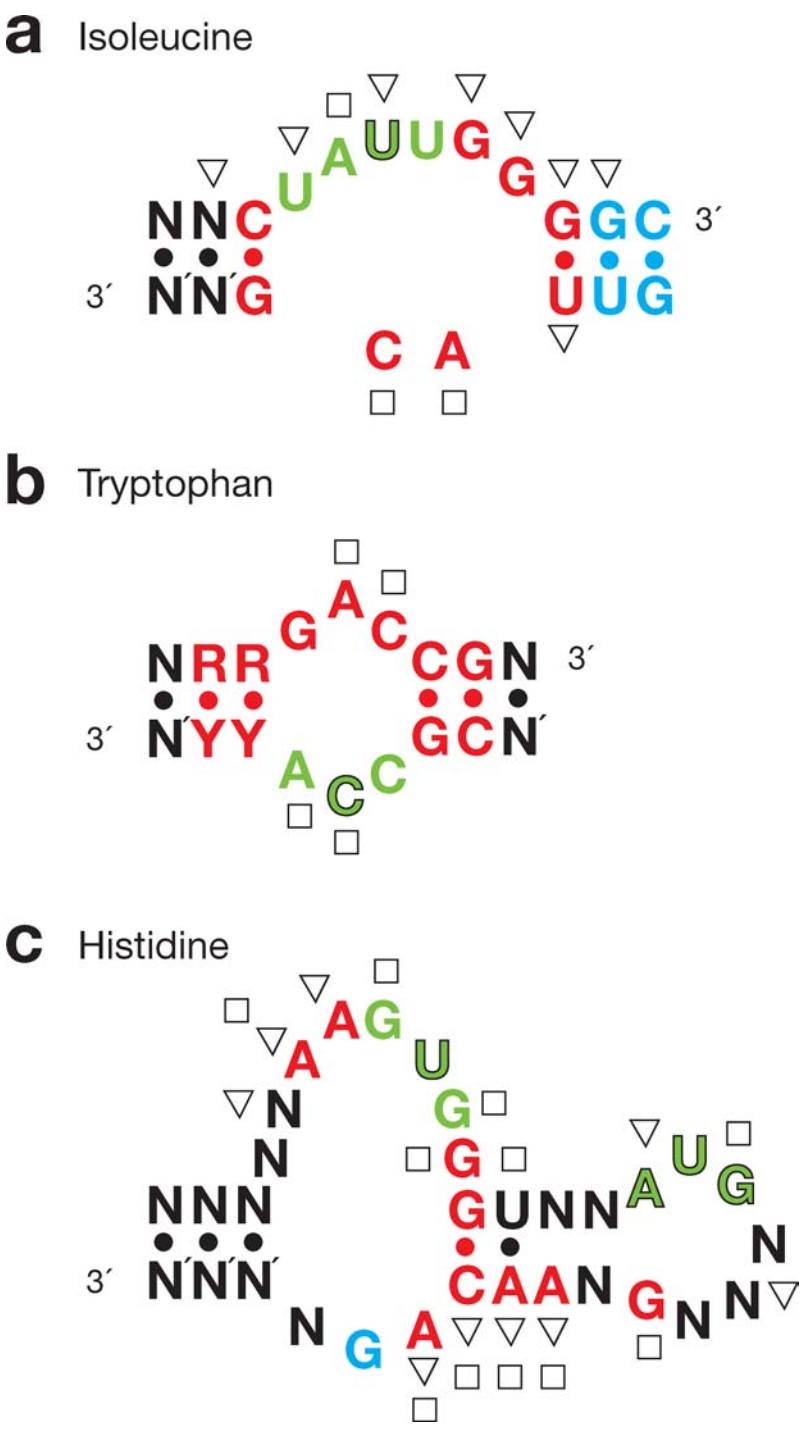
The statistical argument for concentration of triplets from the present code in binding sites is substantial, but it leaves many questions open. In view of the large number of known sites for amino acids, how were the particular ones that gave rise to the code chosen? Unexpectedly, an experimental answer to this question can be constructed. The simplest RNA site for binding each of three amino acids is known from experiment, and this simplest site (containing the fewest nucleotides) contains coding triplets in each case. Thus, the most easily isolated aptamers were likely chosen for the biological code, using a logic that one of us has called exiguity (28). To clarify this argument, we review the RNA sites for isoleucine, tryptophan, and histidine.

Isoleucine

Many selections have been directed at isoleucine-binding RNAs. In initial attempts, a simple internal loop with a dissociation constant (K_D) of several hundred μM was independently isolated five times (29). This structure, now isolated independently more than 63 times, along with its conserved sequences and modification interferences, is shown in Figure 1*a*. In the absence of other considerations, the simplest structure, requiring the fixation of the fewest nucleotides, should be the

Figure 1 Amino acid aptamers that are arguably the simplest possible specific, functional RNAs. The abbreviations are N, any nucleotide; N′, nucleotide complementary to N; R, purine; and Y, pyrimidine. All colored nucleotides are conserved. Red indicates 90% to complete conservation; blue, 60% to 80% conservation; and green, conserved coding triplets. Green nucleotides are totally conserved, except when shown as outlined green; isoleucine, 82% of sequences are U; tryptophan, 30% of isolates are C; histidine, 62% of sequences are U. Squares indicate ligand-protected or enhanced chemical reactivities; triangles show chemical modification-interference with ligand binding.

11 Feb 2005 16:55 AR AR261-BI74-07.tex XMLPublish^{SM}(2004/02/24) P1: JRX
AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119



11 Feb 2005 16:55 AR AR261-BI74-07.tex XMLPublish^{SM}(2004/02/24) P1: JRX
AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119

most frequent outcome of a selection (30, 31). This isoleucine motif was the most frequent specific isolate in its initial selection and is therefore a candidate for simplest solution to the problem of specifically binding isoleucine.

However, during the long cyclic course of selection, there are other possible ways to gain prominence, so the simplest structure may not be the most frequent (32). For example, superior amplification, rather than function, can be decisive (33). Therefore, a procedure comprising selection with decreasing numbers of randomized nucleotides until selection fails to yield a successful isolate was devised. Such a series of selections steadily increases the selective pressure for small motifs, and its ultimate failure suggests that not enough space has been provided for the existence of the function. The same isoleucine-binding internal loop dominated selections having 50, 26, and 22 nucleotides, but it (and any alternative) failed to appear using 16 consecutive randomized nucleotides (32). Thus, this motif is probably the simplest isoleucine-binding RNA and should be prominent in any selection for isoleucine affinity. Notably, this loop structure features both a codon and an anticodon in the sequence UAUU: This core motif is both highly conserved and directly implicated in function by interference assays.

Tryptophan

In unpublished selections (I. Majerfeld and M. Yarus, in preparation) for binding to several amino acids, the RNA of Figure 1*b* was isolated by selecting for specific L-tryptophan affinity. It was recovered from 47 independent parental sequences of various sizes. This motif was observed in populations selected from pools with 70, 60, 40, and 20 randomized positions, and the motif was the predominant isolate at all sizes. When the randomized tract was reduced to 17 nucleotides, this simple structure, a symmetrical 3-nucleotide internal loop flanked by normal base pairs, did not recur (and no other specific free tryptophan aptamer appeared). Thus, this motif meets our most rigorous criteria indicating that it is the simplest structure capable of binding carboxyl-immobilized tryptophan and being eluted with free L-tryptophan.

In the loop, there is one position, tolerating either pyrimidine (Figure 1*b*) that was 20% to 38% C, without an apparent pattern in the different selections. This is a part of the anticodon for tryptophan, 5′ CCA 3′. The chemical data show that this sequence is not only conserved but protected by tryptophan binding, so it is a part of the amino acid–binding site.

Histidine

As for isoleucine and tryptophan above, there is one easily observable solution to the problem of binding protonated L-histidine (Figure 1*c*; I. Majerfeld, D. Puthenvedu, and M. Yarus, in preparation). This was the only solution observed because 82% of the sequences (110 isolates) in the selection from 70 randomized nucleotides constituted 54 independent recurrences of the same site (Figure 1*c*). The other 18% of the sequences were unique and yielded only isolates that neither

11 Feb 2005 16:55 AR AR261-BI74-07.tex XMLPublish^{SM}(2004/02/24) P1: JRX
AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119

bound the column matrix nor eluted with histidine, so probably they are uninteresting in this context. Fragments that include the conserved sequences are functional, establishing that the structure shown is sufficient for activity with a K_D of about 10 μM. Chemical probing of 13 different examples shows that the conserved 5′ GUG 3′ and 5′ AUG 3′ sequences, anticodons for histidine, contain nucleotides protected by bound amino acid. Although the data are less complete for the histidine site than for isoleucine and tryptophan above, 54 independent recurrences of this site with no alternatives probably means that no simpler histidine site exists.

Coding Triplets and Binding Sites are Both Physically and Statistically Associated

For isoleucine, tryptophan, and histidine (the three amino acids for which there is substantial evidence on the simplest aptamer), the simplest, most likely amino acid–binding structure contains functional triplets. Thus, in varied circumstances, selection for amino acid affinity would be expected to produce coding triplets in close approximation to the bound amino acids.

Stereochemical and Adaptive Explanations are Compatible

With a strong effect on the code's structure attributable to stereochemistry, it might be supposed that other factors, such as adaptation and coevolution, would be relegated to lesser roles. However, this is not necessarily true at all; we now show that even a large stereochemical component is compatible with the idea that the code underwent adaptive evolution for error minimization.

In order to test this idea, we produced random variants of the canonical genetic code by fixing 1 to 18 of the actual amino acid assignments to those of the standard genetic code and permuting the amino acid assignments of the rest (Figure 2) (27). We optimized the unfixed remainder of the code by computation, using the Great Deluge algorithm (34) to permute codes across coding blocks in such a way that average change in hydrophobicity from a single mutation was minimized. The results show, at the right in Figure 2, that the real genetic code is highly optimized compared to randomized codes; this reproduces Freeland & Hurst (6). The real genetic code is not optimal on the basis of this criterion but is highly optimized (lower line; Figure 2). In particular, at the levels of preexisting stereochemical assignment we find, it is plausible to suggest that the actual genetic code had a preexisting nucleus of stereochemical assignments, and the remainder of assignments were optimized by subsequent selection for protein function to reach the high degree of optimization found in the real code. Bear in mind that current findings of seven stereochemical assignments out of eight do not constrain the final stereochemical fraction of the code very tightly; at least (7/20=) 35% of modern codons were apparently assigned stereochemically, but the upper experimental limit is presently (19/20=) 95% stereochemical.

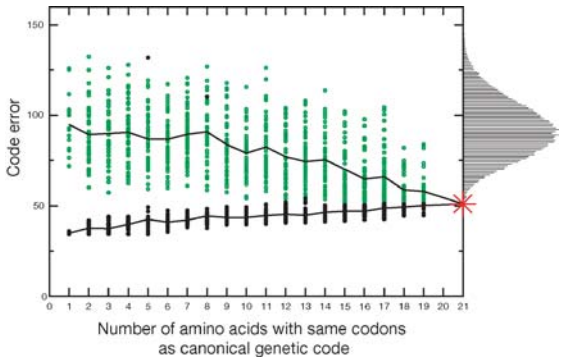


Figure 2 Even a mostly stereochemical code allows an adaptive result. This figure shows as open circles the error values of 50 random genetic codes (y-axis) that share the codons for up to n amino acids with the actual genetic code (x-axis). The actual genetic code is marked with a * at the right, with 21 amino acids shared (the 20 canonical amino acids plus termination). For each random code, we applied a nonlinear optimizer, the Great Deluge algorithm (34), to find the best possible code with that number of shared codons (*solid circles*). The histogram on the right shows the distribution of error values for 100,000 completely randomized codes generated by the same model (amino acid permutation among existing amino acid blocs). The means of the random and optimized codes are plotted as solid lines.

Implications of These Results for the Process of Code Evolution

We conclude that some stereochemical assignments survive to the present. These coding triplets, codons, and anticodons appear to have escaped from primordial amino acid–binding structures made of RNA (or a very similar molecule). This is the central premise of the escaped triplet theory for the code's origin. What was the function of these amino acid–binding structures? One possibility is that they were templates on which activated amino acids were directly ordered for polymerization. The possibility that the primordial template for coded peptide assembly was an RNA-like oligomer of ordered amino acid–binding sites is called the DRT (Direct RNA Template) hypothesis (35). DRT is compared with three alternative roles for primordial amino acid–binding sites in Knight & Landweber (36).

Conversely, some coding assignments may have been made by other processes, although this notion depends on a negative result (cf. Table 1). In particular, the adaptive hypothesis, which suggests that optimization has improved the code's error-minimizing properties, is still plausible given a broad range of amino acids whose coding was predetermined by stereochemistry (Figure 2). In addition, coevolution may have played a role. However, once it is accepted that the code's history records more than one effect (for example, it is stereochemical but evolved along pathways also), uncertainty about the specific pathways to be chosen for partial coevolution makes it difficult to estimate a fraction of the

11 Feb 2005 16:55 AR AR261-BI74-07.tex XMLPublish^{SM}(2004/02/24) P1: JRX
AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119

code potentially attributable to coevolutionary codon assignment (see Coevolution above).

Because associations between coding triplets and RNA-binding sites have no meaning until the appearance of an RNA-like molecule, the fact that binding sites selected from random RNA sequences recapitulate features of the genetic code suggests that these primordial assignments occurred in an RNA world, rather than at a hypothetical earlier stage of evolution. This idea does not rule out the possibility that short peptides were important in metabolism before the origin of RNA but suggests that such peptides were not encoded by a mechanism resembling the present translation system. Because an RNA world itself implies a fairly complex metabolic repertoire (37), the initial genetic code would not necessarily have been limited to amino acids produced by prebiotic mechanisms. To put this another way, the properties of RNA in the translation system make clear that genetic coding was introduced by our immediate biological progenitors (23, 24), not by ancient ancestors near the origin of life.

This idea is reemphasized here in the finding that amino acids with complex biosyntheses, like tryptophan, histidine, and arginine, were included in the presumably primordial stereochemical core of the genetic code. Amino acids, such as histidine, produced by pathways related to nucleotide metabolism, but not found in spark-tube experiments, are in fact excellent candidates for inclusion because they interact well with the RNA that must be present. This underscores the notion that the origin of the genetic code is a different and later event than the origin of life; it employed amino acids that lie at the end of what must have been highly evolved anabolic pathways. These pathways must have evolved before the code, and ribozymes are obvious candidates for their catalysis.

THE BIOCHEMICAL AND INTELLECTUAL SETTING FOR THE ORIGIN OF THE GENETIC CODE

We finish with two short sections, first about the biochemical and then the logical setting for studies of the code's origin. Having raised the first topic just above, we summarize recent findings that bear on the biochemical environment in which the genetic code might have appeared.

Given activated nucleotides from pre-RNA world activities, montmorillonite clay catalyzes assembly of RNA oligomers, polymerizing to 35- to 40-mers in 24 h (38). The polymerization process may also be accelerated when dilute solutions are frozen to provide liquid regions of concentrated ions and nucleotides within an ice phase (39). Curiously, these frozen conditions also lessen the sequence requirements and emphasize the ligation reactions of the hairpin ribozyme (40). Thus, it appears that ligating ribozymes might convert shorter synthetic oligomers to populations containing longer RNAs at low temperatures that stabilize product RNAs. Longer RNAs are particularly important because there is evidence that more capable aptamers and ribozymes will usually require longer folded sequences (41).

11 Feb 2005 16:55 AR AR261-BI74-07.tex XMLPublish^{SM}(2004/02/24) P1: JRX
AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119

RNA is unstable and has complex, necessarily rare precursors, so the accumulation of sufficient material to initiate the metabolism implied by an RNA world has always been a salient problem. Therefore, it is of interest that these threshold amounts may be much less than intuition suggests. The statistical advantages of active sites divided into pieces (30) or modules mean they can be more frequent among random sequence RNAs, especially if the separate sequence modules are of about the same size (31). The numerical effect of these ideas raises the frequencies of required RNA sequences by orders of magnitude; populations of zeptomoles (6×10^2 s) or attomoles (6×10^5 s) of 80-mer RNA sequences may be enough to find the required sequences for some known ribozymes and aptamers. However, these earlier calculations underestimate the number of required random sequences in a large and a small way (41a). The small underestimation is due to folding; only a fraction of existing sequences will fold. This fraction has been estimated by thermodynamic folding corrections. The more serious underestimation is that the difficulties of finding paired sequences around required modules were not directly calculated in earlier estimates (30, 31). Heuristic methods for estimating pairing probability can be used to show that real motifs like the isoleucine aptamer (29, 32) and the hammerhead ribozyme (41b), corrected for folding, should appear among several times 10^9 to 10^{10} randomized RNAs. This corresponds to nanograms of 100-mer RNAs, which we now estimate as the minimal pool size required to initiate an RNA world (30). These quantities are 10,000- to 100,000-fold smaller than those used in modern selection experiments. Therefore, an RNA world is more accessible than might have been thought.

In addition to the quantitative problems of RNA synthesis, new experiments also make clear that the qualitative aspects of active RNA synthesis are subject to fewer constraints than once thought. For example, both strands of a completely complementary hybrid helix can be nucleic acid enzymes, as in the combination of a DNA ribonuclease strand and a hairpin RNAse strand (42). Furthermore, a moderately active ligase ribozyme can be built using just two nucleotides, 2,6-diaminopurine and uracil (43).

Given sufficiently large RNA sequence populations to search for new activities, it is clear that the search would not find any obvious limit. In the next section, we talk about the kind of activities that are the most apt for biology and the ribocyte, but the recent selection of RNAs that speed formation of palladium metal particles (44) or that capture long UV light and reverse UV base dimer photodamage (45) both introduce novel RNA capabilities.

Studying Molecular Evolution via Selection

Finally, how can we learn about evolution in an RNA world? The difficulty is one of inaccessibility. That is, a surprising fraction of everyday scientific concerns are entirely out of reach of direct experimentation. Without the ability to do experiments, we may ask how investigators make progress. Most will agree that we do know some reliable things about (for example) cosmology, geophysics, climatology, ecology, and other historical disciplines. However, a particular remoteness

11 Feb 2005 16:55 AR AR261-BI74-07.tex XMLPublish^{SM}(2004/02/24) P1: JRX
AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119

afflicts all evolutionary inquiries. Where does scientific knowledge come from in experimentally inaccessible areas?

An answer to many such questions came from Thomas Bayes, an eighteenth-century British cleric who devised the following:

$$P(H|D) = P(H)\frac{P(D|H)}{P(D)},$$

where H is a hypothesis, D is data, P is probability, and the vertical bar is read "given." Even without analysis, it is easy to appreciate the subject under discussion: Bayes' theorem specifies how we can revise the probability of a hypothesis, $P(H)$ when we get new data, $P(H|D)$. The factor by which a good hypothesis is made more likely is the factor by which the likelihood of the data if the hypothesis is true, $P(D|H)$, exceeds the overall probability of observing the data in light of all hypotheses, $P(D)$. Bayes' theorem embodies the commonsense idea that unique observations, unlikely in general but expected on a particular hypothesis, supply good confirmation. However, Bayes' has the virtue over common sense of being not only true but also quantitative, general, and optimal.

Further, if there are several predictions and several independent kinds of evidence, they can be used together to update the probability of a hypothesis. So, for independent experiments giving data D_1, D_2,..., D_j, each updated probability is the starting point for updating by subsequent evidence using

$$P(H|D_1D_2\ldots D_j) = P(H)\frac{P(D_1|H)}{P(D_1)} \times \frac{P(D_2|H)}{P(D_2)} \times \ldots \times \frac{P(D_j|H)}{P(D_j)}.$$

Multiple confirmed predictions, therefore, multiply their effects on the plausibility of the initial idea, $P(H)$. Results 10-fold more likely on a given hypothesis than in general raise the probability of the hypothesis 10-fold; two such independent experiments raise it 100-fold, and so on.

In fact, Bayesian inquiry and confirmation can seem so effective that perhaps we should worry that it is too good to be true; that is, perhaps we should worry about convergence to $P(H|D) = 1$. For a true hypothesis H_i, one may envision any number of supportive experiments starting at any prior credibility $P(H_i)$; so why does its probability never exceed 1?

$$P(D) = \sum_i P(H_i) \times P(D|H_i),$$

where $P(D)$ is expanded to show how it can be summed over an exhaustive set of i possible hypotheses: H_1, H_2,..., H_i. As the probability of a true hypothesis $P(H_i)$ grows, the term $P(H_i)\,P(D|H_i)$ relating to the true hypothesis also grows. Near the limit of confirmation, $P(H_i)\,P(D|H_i)$ is both the leading term in the denominator and the complete numerator for $P(H_i|D)$, so $P(H_i|D)$ increases to 1.0 precisely.

$$P(H_i|D) = \frac{P(H_i) \times P(D|H_i)}{\sum_i P(H_i) \times P(D|H_i)} \rightarrow 1.0.$$

11 Feb 2005 16:55 AR AR261-BI74-07.tex XMLPublish^{SM}(2004/02/24) P1: JRX
AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119

Getting data favorable to a progressively confirmed hypothesis is less and less surprising, so experiments increase the probability of H_i by a smaller and smaller factor.

The application of Bayes' theorem to scientific questions that allow no direct experimentation can now be summarized. Any idea that makes predictions can be tested and perhaps confirmed, whether the initial subject can be apprehended and brought into the laboratory for questioning, or not. As long as the predictions themselves are currently apt, even hypothetical events unimaginably remote in time or space can be investigated.

Consider as an example the hypothesis that the extinctions at the K-T (Cretaceous-Tertiary) boundary were caused by an extraterrestrial impact (46). The differentiation of the Earth's core depleted the crust of heavy elements, including Pt, Os, Ir, and Rh, with respect to average solar system material. Therefore, the finding of a 30- to100-fold iridium spike in K-T boundary clays suggested a sudden influx of extraterrestrial material around the time of the K-T extinction (which ablated about half of terrestrial genera). That is, because iridium (Ir) would have been relatively elevated in extraterrestrial impactors, $P(D|H)$ for excess Ir is high given the hypothesis that an such an impactor spread terrestrial and bolide material all over Earth. A correct hypothesis attracts other confirming results, adding other data whose probabilities were unusually high only in the light of the hypothesis followed up by the Ir hint. For example, along with Ir excesses were quartz grains that recorded an impact shock (47) and glassy tektites conceivably formed using impact energy for fusion (48). Sometimes, around the Caribbean basin, all these occur together along with indications of powerful tidal activity (49). And the ultimate of these $P(D|H)$ in one sense was the identification of Earth's largest impact crater, the 180 km diameter Chicxulub crater in the Yucatan (50), which indeed dates from around the K-T boundary.

Several aspects of this argument are especially notable for our purposes. The K-T extinction is crucial in any account of the history of life on Earth but is distant in time and scale, securely out of range of any conceivable experiment. Exceptional Ir abundance is plausible for an extraterrestrial encounter [$P(D|H)$ high] but implausible for earthly crustal sources [$P(D)$ low]. This elemental observation attracts other evidence that suggests a rare, large, devastating impact in the Caribbean. For example, an exceptional [low $P(D)$, but high $P(D|H)$] remnant crater exists in the region implicated. Thus, a convincing web of evidence with each item improbable, except in the light of the hypothesis of an exploding impactor, when taken together lead to a persuasive conclusion about an event that can never be perceived, much less experimentally manipulated.

The K-T example logically parallels many other scientific arguments. In particular, it parallels varied evolutionary arguments and, more particularly, the argument for some kind of RNA world with an origin of the genetic code therein. All that is necessary is to recast the hypothetical ancient event as a set of current predictions.

There are, on the one hand, many activities predicted for cells that use RNA as their primary macromolecule (ribocytes). In order to diversify into something

11 Feb 2005 16:55 AR AR261-BI74-07.tex XMLPublish^{SM}(2004/02/24) P1: JRX
AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119

sufficiently complex to be recognizable as biological, a ribocyte must probably have access to Darwinian evolution and therefore replicate. This requires storage and duplication of biological information, for which there is no empirical example outside of nucleic acids [although other possibilities are sometimes asserted (51)]. In fact, this prediction can be confirmed by the isolation of a rudimentary RNA replicase ribozyme (52, 53). To progress to nucleoprotein cells from the ribocyte era, the RNA cell must devise translation. Therefore, the four essential reactions of translation must be part of the RNA repertoire. This prediction has been confirmed by the discovery of amino acid–activating (54), –aminoacylating (55, 56), peptidyl-transfer RNAs (57), as well as RNAs that retain a residue of the genetic code (23) (see above). Ribocytes would require modulation of the permeability of bounding membranes, and membrane binding and transport activity have also been demonstrated in RNAs (58). A plausible RNA metabolism must exist, and this prediction can be partially confirmed by the RNA-catalyzed construction of ribonucleotide cofactor-using RNA catalysts (59, 60). With each successful step in this progression, Bayes' theorem tells us that we must increase our estimate of the probability of the ribocyte.

More specifically, this review is concerned with one of the steps in the emergence of template-guided translation and with the genetic code. Here, a hint from a putative molecular fossil, the *Tetrahymena* arginine site (61), has been expanded into a now-extensive study of 49-amino acid–binding sites for 8 of the standard 20-biological side chains. The findings are that, in this population of sites, both codons and particularly anticodons are improbably frequent in the sites of amino acid aptamers as a whole. This conclusion relies on seven cases of a total of eight amino acids: isoleucine, arginine, tryptophan, histidine, leucine, phenylalanine, and tyrosine. In contrast, only glutamine shows no significant concentration of anticodon and codon triplets in its known binding sites (Table 1). Therefore we presently estimate that 7/8 or 88% of the amino acids had stereochemical codon assignments, derived from the exploitation of sequences from amino acid–binding sites [e.g., (35)], and these assignments have survived in the prevalent coding table. This escaped triplet theory can be carried another step for isoleucine, tryptophan and histidine, in which there is reason to believe that the simplest (and therefore often most prevalent) active binding site structures have been isolated. These simplest sites do contain cognate codons and anticodons, which therefore are likely to recur in connection with any selection, present or primordial, for which specific amino acid affinity was demanded from RNA-like molecules.

Thus, we argue that an initial hint about the origin of the genetic code has attracted independent confirmatory observations. In fact, it seems to us that a proper set of conclusions would be stronger than so far stated. It is customary to observe that some newly selected RNA capability supports the existence of an RNA world or a genetic code based on RNA chemistry. In fact, there are now many wo/man years of results, indicating that RNA does many biologically relevant related things it is not known to do in current cellular life. These observations join to make a potent argument that such potentialities were once used together. We are

11 Feb 2005 16:55 AR AR261-BI74-07.tex XMLPublish[SM](2004/02/24) P1: JRX
AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119

not saying that this sort of indirect Bayesian investigation is, in any sense, better than straightforward experimentation. On the contrary, it is usually less efficient because it must be less focused and tolerate broader gaps. Nonetheless, an indirect Bayesian approach produces scientific knowledge (as enhanced probabilities for specific ideas) in the same sense that experimental studies of more accessible systems do, and this point seems rarely appreciated.

ACKNOWLEDGMENTS

We thank Carol Cleland and Massimo Di Giulio, as well as members of the Yarus group, for discussions about draft manuscripts. Preparation of this review was supported by NIH research grant GM 48080 and the NASA Center for Astrobiology NCC2-1052.

The *Annual Review of Biochemistry* is online at http://biochem.annualreviews.org

LITERATURE CITED

1. Jukes TH, Osawa S. 1993. *Comp. Biochem. Physiol. B* 106:489–94
2. Woese CR. 2002. *Proc. Natl. Acad. Sci. USA* 99:8742–47
3. Woese CR. 1965. *Proc. Natl. Acad. Sci. USA* 54:1546–52
4. Knight RD, Freeland SJ, Landweber LF. 1999. *Trends Biochem. Sci.* 24:241–47
5. Sonneborn TM. 1965. In *Evolving Genes and Proteins*, ed. V Bryson, HJ Vogel, pp. 377–97. New York: Academic
6. Freeland SJ, Hurst LD. 1998. *J. Mol. Evol.* 47:238–48
7. Freeland SJ, Wu T, Keulmann N. 2003. *Orig. Life Evol. Biosph.* 33:457–77
8. Ardell DH, Sella G. 2002. *Philos. Trans. R. Soc. London Ser. B* 357:1625–42
9. Wong JT-F. 1975. *Proc. Natl. Acad. Sci. USA* 72:1909–12
10. Amirnovin R. 1997. *J. Mol. Evol.* 44:473–76
11. Di Giulio M. 1999. *J. Mol. Evol.* 48:253–55
12. Ronneberg TA, Landweber LF, Freeland SJ. 2000. *Proc. Natl. Acad. Sci. USA* 97:13690–95
13. Di Giulio M. 2001. *J. Mol. Evol.* 53:724–32
14. Wilcox M, Nirenberg MW. 1968. *Proc. Natl. Acad. Sci. USA* 61:229–36
15. Curnow AW, Ibba M, Soll D. 1996. *Nature* 382:589–90
16. Crick FHC. 1968. *J. Mol. Biol.* 38:367–79
17. Schultz DW, Yarus M. 1996. *J. Mol. Evol.* 42:597–601
18. Knight RD, Landweber LF, Yarus M. 2001. *J. Mol. Evol.* 53:299–313
19. Santos MA, Moura G, Massey SE, Tuite MF. 2004. *Trends Genet.* 20:95–102
20. Orgel LE. 1968. *J. Mol. Biol.* 38:381–93
21. Joyce GF. 2002. *Nature* 418:214–21
22. Gesteland RF, Cech T, Atkins JF, eds. 1999. *The RNA World*. New York: Cold Spring Harbor Lab. Press. 709 pp.
23. Yarus M. 2001. *Cold Spring Harbor Symp. Quant. Biol.* 66:207–15
24. Yarus M. 2002. *Annu. Rev. Genet.* 36:125–51
25. Knight RD, Landweber LF, Yarus M. 2003. In *Translation Mechanisms*, ed. J Lapointe, L Brakier-Gingras, pp. 115–28. New York: Kluwer Acad.

11 Feb 2005 16:55 AR AR261-BI74-07.tex XMLPublish^{SM}(2004/02/24) P1: JRX
AR REVIEWS IN ADVANCE10.1146/annurev.biochem.74.082803.133119

26. Ellington AD, Khrapov M, Shaw CA. 2000. *RNA* 6:485–98
27. Caporaso JG, Yarus M, Knight RD. 2005. *J. Mol. Evol*. In press
28. Yarus M, Welch M. 2000. *Chem. Biol.* 7:R187–90
29. Majerfeld I, Yarus M. 1998. *RNA* 4:471–78
30. Yarus M, Knight RD. 2004. In *The Genetic Code and the Origin of Life*, ed. LR Pouplana, pp. 75–91. Georgetown, TX: Landes Biosci.
31. Knight R, Yarus M. 2003. *RNA* 9:218–30
32. Lozupone C, Changayil S, Majerfeld I, Yarus M. 2003. *RNA* 9:1315–22
33. Coleman TM, Huang F. 2002. *Chem. Biol.* 9:1227–36
34. Dueck G. 1993. *J. Comput. Phys.* 104:86–92
35. Yarus M. 1998. *J. Mol. Evol.* 47:109–17
36. Knight RD, Landweber LF. 2000. *RNA* 6:499–510
37. Yarus M. 1999. *Curr. Opin. Chem. Biol.* 3:260–67
38. Huang WH, Ferris JP. 2003. *Chem. Commun.* (12):1458–59
39. Monnard P, Kanavarioti A, Deamer D. 2003. *J. Am. Chem. Soc.* 125:13734–40
40. Vlassov AV, Johnston BH, Landweber LF, Kazakov SA. 2004. *Nucleic Acids Res.* 32:2966–74
41. Carothers JM, Oestreich SC, Davis JH, Szostak JW. 2004. *J. Am. Chem. Soc.* 126:5130–37
41a. Knight RD, DeSterk H, Markel R, Smit S, Oshmyansky A, Yarus M. 2005. *Nucleic Acids Res*. In press
41b. Salehi-Ashtiani K, Szostak JW. 2001. *Nature* 414:82–84
42. Kuhns ST, Joyce GF. 2003. *J. Mol. Evol.* 56:711–17
43. Reader JS, Joyce GF. 2002. *Nature* 420: 841–44
44. Gugliotti LA, Feldheim DL, Eaton BE. 2004. *Science* 304:850–52
45. Chinnapen DJF, Sen D. 2004. *Proc. Natl. Acad. Sci. USA.* 101:65–69
46. Alvarez LW, Alvarez W, Asaro F, Michel HV. 1980. *Science* 208:1095–108
47. Bohor B, Foord EE, Modreski PJ, Triplehorn DM. 1984. *Science* 224:867–69
48. Sigurdsson H, D'Hondt S, Arthur MA, Bralower TJ, Zachos JC, et al. 1991. *Nature* 349:482–87
49. Smit J, Montanari A, Swinburne NH, Alvarez W, Hildebrand AR, et al. 1992. *Geology* 20:99–103
50. Hildebrand AR, Penfield GT, Kring DA, Pilkington M, Camargo ZA, et al. 1991. *Geology* 19:867–71
51. Cairns-Smith AG. 1982. *Genetic Takeover and the Mineral Origins of Life*. Cambridge, UK: Cambridge Univ. Press
52. Johnston WK, Unrau PJ, Lawrence MS, Glasner ME, Bartel DP. 2001. *Science* 292:1319–25
53. Lawrence MS, Bartel DP. 2003. *Biochemistry* 42:8748–55
54. Kumar RK, Yarus M. 2001. *Biochemistry* 40:6998–7004
55. Illangasekare M, Sanchez G, Nickles T, Yarus M. 1995. *Science* 267:643–47
56. Murakami H, Saito H, Suga H. 2003. *Chem. Biol.* 10:655–62
57. Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. 2000. *Science* 289:920–30
58. Janas T, Yarus M. 2003. *RNA* 9:1353–61
59. Jadhav VR, Yarus M. 2002. *Biochemistry* 41:723–29
60. Tsukiji S, Pattnaik SB, Suga H. 2004. *J. Am. Chem. Soc.* 126:5044–45
61. Yarus M, Christian EL. 1989. *Nature* 342:349–50
62. Knight RD, Landweber LF. 1998. *Chem. Biol.* 5:R215–20
63. Illangasekare M, Yarus M. 2002. *J. Mol. Evol.* 54:298–311
64. Burgstaller P, Kochoyan M, Famulok M. 1995. *Nucleic Acids Res.* 23:4769–76
65. Calnan BJ, Tidor B, Biancalana S, Hudson D, Frankel AD. 1991. *Science* 252:1167–71
66. Connell GJ, Illangsekare M, Yarus M. 1993. *Biochemistry* 32:5497–502

67. Geiger A, Burgstaller P, von der Eltz H, Roeder A, Famulok M. 1996. *Nucleic Acids Res.* 24:1029–36
68. Connell GJ, Yarus M. 1994. *Science* 264:1137–41
69. Tao J, Frankel AD. 1996. *Biochemistry* 35:2229–38
70. Yang Y, Kochoyan M, Burgstaller P, Westhof E, Famulok M. 1996. *Science* 272:1343–46
71. Yarus M. 1993. In *The RNA World*, ed. RF Gesteland, JF Atkins, pp. 205–17. New York: Cold Spring Harbor Lab. Press
72. Mannironi C, Scerch C, Fruscoloni P, Tocchini-Valentini GP. 2000. *RNA* 6:520–27