OPEN ACCESS

BIO-complexity

Critical Review

Can the Origin of the Genetic Code Be Explained by Direct RNA Templating?

Stephen C. Meyer* and Paul A. Nelson

Discovery Institute, Seattle, WA, USA

Abstract

Motivated by the RNA world hypothesis, Michael Yarus and colleagues have proposed a model for the origin of the 'universal' genetic code, in which RNA aptamers directly template amino acids for protein assembly. Yarus *et al*. claim that this "direct RNA templating" (DRT) model provides a stereochemical basis for the origin of the code, as shown by the higher-than-expected frequency of cognate coding triplets in aptamer amino acid-binding sites. However, the DRT model suffers from several defects. These include the selective use of data, incorrect null models, a weak signal even from positive results, an implausible geometry for the primordial RNA template (in relation to the universally-conserved structures of modern ribosomes), and unsupported assumptions about the pre-biotic availability of amino acids. Although Yarus *et al.* claim that the DRT model undermines an intelligent design explanation for the origin of the genetic code, the model's many shortcomings in fact illustrate the insufficiency of undirected chemistry to construct the semantic system represented by the code we see today.

Cite as: Meyer SC, Nelson PA (2011) Can the origin of the genetic code be explained by direct DNA templating? BIO-Complexity 2011(2): 1-10. **doi:**10.5048/BIO-C.2011.2

Editor: Peter Imming

Received: March 23, 2011; **Accepted:** August 5, 2011; **Published:** August 24, 2011

Notes: A *Critique* of this paper, when available, will be assigned **doi:**10.5048/BIO-C.2011.2.c.

* Email: stevemeyer@discovery.org

INTRODUCTION

In the four decades following its elucidation in 1966 by Nirenberg, Khorana and Holley [1], the "universal" genetic code has provided the target for a wide array of hypotheses concerning its evolutionary origin. A long library shelf's worth of hypotheses (some of them bare speculations) have accumulated about the origin of the code. In their recent comprehensive review of these hypotheses, Koonin and Novozhilov call the problem of the origin of the code a "universal enigma"[2]. They explain that "despite extensive and, in many cases, elaborate attempts to model code optimization, ingenious theorizing along the lines of the coevolution theory, and considerable experimentation, very little definitive progress has been made" [2].

Some observers have claimed, however, that recent work has broken this long-standing impasse. They cite the experiments of Michael Yarus and colleagues at the University of Colorado to show that the origin of the genetic code can be explained as the result of stereochemical affinities between RNA triplets and the corresponding (cognate) amino acids with which they are associated in the code. Indeed, Yarus himself has asserted that his work not only demonstrates how the code evolved, but also undermines a key claim of the theory of intelligent design, by showing that specified complexity can arise by purely natural processes [3].

Of the handful of hypotheses on the origin of the code that claim experimental support, Yarus and colleagues' is arguably the best-articulated. They argue that RNA provides a direct (stereochemical) template for the specific binding of proteinaceous amino acids, and that chemical affinities between RNA triplets and specific amino acids initially formed the basis of the "universal" genetic code. Moreover, Yarus *et al.* have assembled a significant body of novel experimental data, which they argue support their hypothesis [4-11].

In this review, we shall evaluate their Direct RNA Template stereochemical model (hereafter, the DRT model) for the origin of the genetic code. We do this not only because various commentators have claimed that the DRT model refutes a key argument for intelligent design, but also because the model has obvious relevance to the perennial problem of the origin of life, and the closely related problem of the origin of biological information.

ANALYSIS

The genetic code as we find it today

To judge the significance of the DRT model, laid out most fully by Yarus *et al.* in their 2009 review [11], we need first to describe what needs to be explained in more detail.

The genetic code as we observe it today is a semantic (symbol-based) relation between (a) amino acids, the building blocks of proteins, and (b) codons, the three-nucleotide units in messenger RNA specifying the identity and order of different amino acids in protein assembly (Fig. 1).

The actual physical mediators of the code, however, are transfer RNAs (tRNAs) that, after being charged with their specific amino acids by enzymes known as aminoacyl transfer RNA synthetases (aaRSs), present the amino acids for peptide bond formation in the peptidyl-transferase (P) site of the ribosome, the molecular machine that constructs proteins.

The secondary structure of a typical tRNA (Fig. 2) reveals the coding (semantic) relations that Yarus *et al.* [11] are trying to obtain from chemistry alone – a quest Yockey has compared to latter-day alchemy [12].

At the end of its 3' arm, the tRNA binds its cognate amino acid via the universally conserved CCA sequence. Some distance away—about 70 Å—in loop 2, at the other end of the inverted cloverleaf, the anticodon recognizes the corresponding codon in the mRNA strand. (The familiar 'cloverleaf' shape represents only the secondary structure of tRNA; its three-dimensional form more closely resembles an "L" shape, with the anticodon at one end and an amino acid at the other.)

Thus, in the current genetic code, there is no direct chemical interaction between codons, anticodons, and amino acids. The anticodon triplet and amino acid are situated at opposite ends of the tRNA: the mRNA codon binds not to the amino acid directly, but rather to the anticodon triplet in loop 2 of the tRNA.

Since all twenty amino acids, when bound to their corresponding tRNA molecules, attach to the same CCA sequence at the end of the 3' arm, the *stereochemical* properties of that nucleotide sequence clearly do not determine which amino

	U	C	A	G
U	UUU Phenylalanine UUC Phenylalanine UUA Leucine UUG Leucine	UCU Serine UCC Serine UCA Serine UCG Serine	UAU Tyrosine UAC Tyrosine UAA Stop UAG Stop	UGU Cysteine UGC Cysteine UGA Stop UGG Tryptophan
C	CUU Leucine CUC Leucine CUA Leucine CUG Leucine	CCU Proline CCC Proline CCA Proline CCG Proline	CAU Histidine CAC Histidine CAA Glutamine CAG Glutamine	CGU Arginine CGC Arginine CGA Arginine CGG Arginine
A	AUU Isoleucine AUC Isoleucine AUA Isoleucine AUG Methionine(Start)	ACU Threonine ACC Threonine ACA Threonine ACG Threonine	AAU Asparagine AAC Asparagine AAA Lysine AAG Lysine	AGU Serine AGC Serine AGA Arginine AGG Arginine
G	GUU Valine GUC Valine GUA Valine GUG Valine	GCU Alanine GCC Alanine GCA Alanine GCG Alanine	GAU Aspartic GAC Aspartic GAA Glutamic GAG Glutamic	GGU Glycine GGC Glycine GGA Glycine GGG Glycine

Figure 1. The 'universal' genetic code. **doi:**10.5048/BIO-C.2011.2.f1

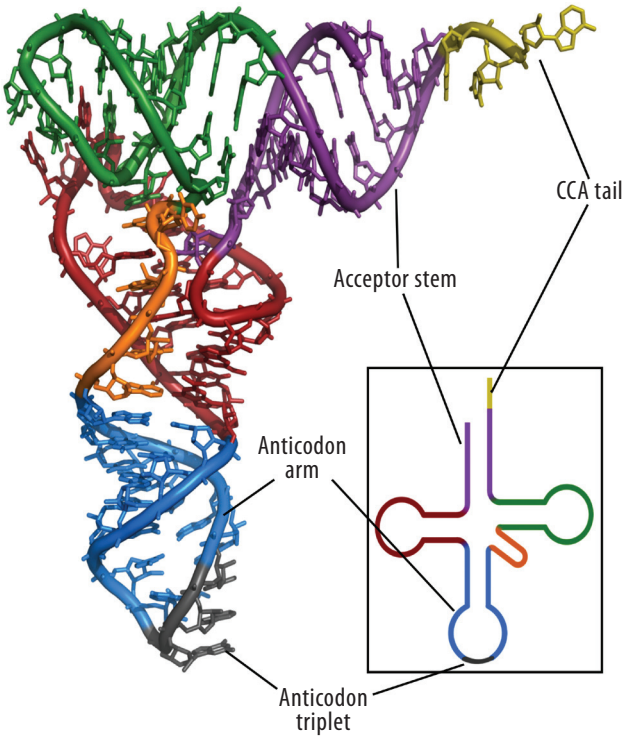

Figure 2. The tertiary structure of phenylalanine tRNA^{phe} from yeast. The boxed area is a schematic diagram of the same tRNA, illustrating its typical cloverleaf secondary structure. Both are colored identically. Adapted from an image by Yikrazuul obtained from Wikimedia Commons (http://commons.wikimedia.org/wiki/File:TRNA-Phe_yeast_1ehz.png) under the Creative Commons Attribution License. **doi:**10.5048/BIO-C.2011.2.f2

acids attach, and which do not. The CCA sequence is indifferent, so to speak, to which amino acids bind to it (just as the sugar-phosphate backbone in DNA is indifferent to which nucleotide bases bind to it). Seen from the bottom-up perspective of chemistry, therefore, the code is *physically* arbitrary.

Nevertheless, tRNAs are *informationally* (i.e., semantically) highly specific: protein assembly and biological function—but not chemistry—demand such specificity. As noted, in the current code, codon-to-amino acid semantic mappings are mediated by tRNAs, but also by the enzymatic action of the twenty separate aminoacyl-tRNA synthetases ("aaRSs"). Most cells use twenty aaRS enzymes, one for each amino acid. Each of these proteins recognizes a specific amino acid and the specific anticodons it binds to within the code. They then bind amino acids to the tRNA that bears the corresponding anticodon.

Thus, instead of the code reducing to a simple set of stereochemical affinities, biochemists have found a functionally interdependent system of highly specific molecules, including mRNA, a suite of tRNAs, and twenty specific aaRS enzymes, each of which is itself constructed from information stored on the very DNA strands that the system as a whole decodes. Attempts to explain one part of the integrated complexity of the gene-expression system, namely the genetic code, by reference to simple chemical affinities lead not to simple rules of chemical attraction, but instead to an integrated system of multiple large molecular components. While this information-transmitting system exploits (i.e., uses) chemistry, it is not *reducible* to direct

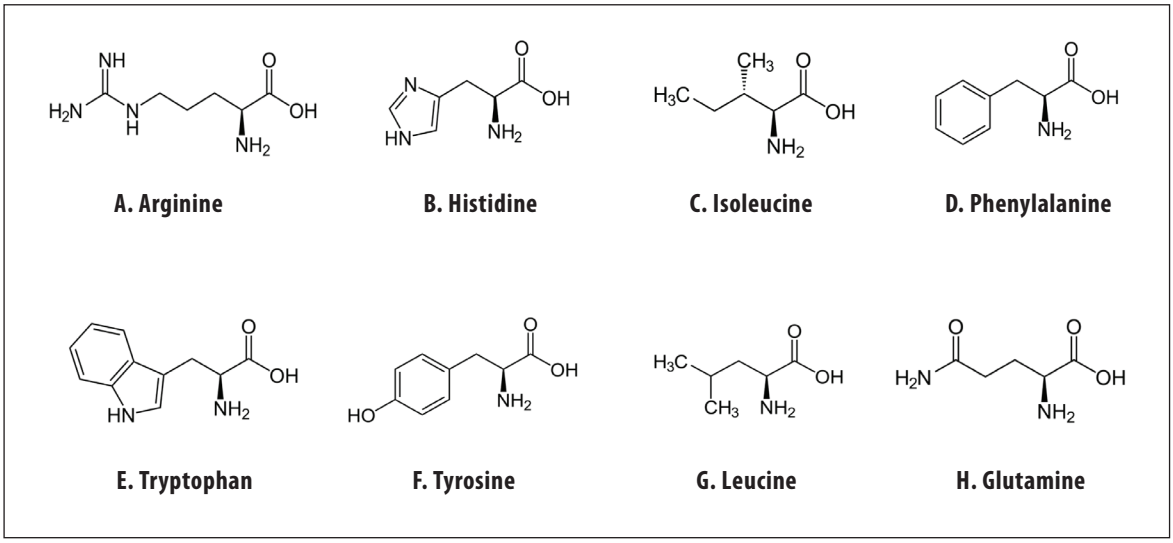

Figure 3. The structures of the eight amino acids assayed by Yarus *et al*. [11]. **doi:**10.5048/BIO-C.2011.2.f3

chemical affinities between codons or anticodons and their cognate amino acids.

But suppose nonetheless that one wanted to derive the semantic or informational properties of this system, which are essential for *biological* function—in particular, the codon-to-amino acid mappings of the genetic code—from chemistry alone, which is indifferent to biological function. That is what Yarus and colleagues set out to do, because their bottom-up, undirected-physics-leads-to-life perspective requires that such a causal story be told, somehow.

Before we summarize their series of experiments, however, we should take note of a plain fact too readily forgotten. As Koonin and Novozhilov observe, a pervasive air of unreality hangs over studies on the origin of the genetic code [2]. Charitably, one might allow that some intractable problems in the pathway from chemicals to life can be leapfrogged, or bracketed, while others are attacked – along the lines of a "let's pretend we will eventually solve that puzzle over there, while we work on this one here" attitude.

But charity can be abused. Synthesizing large pools of RNAs by direct intelligent intervention, for instance, and then claiming that one has modeled undirected abiogenesis on the early Earth, does not count as provisionally bracketing a problem. Thus, we will borrow Koonin and Novozhilov's description, "dubious relevance" [2], to flag various aspects of the experiments of Yarus *et al.*, to keep biological reality reasonably close at hand.

What Yarus *et al.* found and how they interpreted their results

Under the RNA world hypothesis, RNAs known as ribozymes must have once performed essential biochemical functions performed by modern proteins, such as binding specific amino acids for protein assembly (as is carried out today by aaRS enzymes in concert with tRNAs). In the late 1980s, motivated by the RNA world scenario, Yarus *et al.* began to look for RNA-amino acid affinities, because "a translation system made of RNA [i.e., RNA alone, as posited to have existed in the RNA world] must also show chemical selectivity (or there will be no coding)" [11]. Yarus *et al.* posit an early state in which RNA molecules with certain base sequences differentially attract the particular amino acids with which they are associated in the modern genetic code – thus enabling the code to arise directly from stereochemical associations.

Yarus found support for this thesis in an early experiment [4], in which he discovered a differential bonding affinity between the amino acid arginine and RNA bases at the active site in the group I intron of *Tetrahymena*, a ciliated protozoan. He found that arginine inhibits the self-splicing reaction of the group I intron by preferentially binding to sequences containing nucleotides corresponding to arginine codons (AGA, CGA, and AGG). These data led Yarus *et al.* to speculate that the group I intron represented a molecular fossil—showing the *specific* binding of amino acids directly by RNA—which he claimed "developed from an ancient RNA codon-amino acid interaction" [5].

The arginine result was suggestive enough to send Yarus and his colleagues on a search for other amino acid-RNA sequence affinities. To do this, they looked for RNA strands that bound certain amino acids preferentially, from a class of RNA molecules now dubbed *aptamers*. Using the directed evolution method SELEX, in which large pools of random RNA strands are synthesized and then sifted for particular functions (in this case, amino acid binding), Yarus *et al.* [11] generated and characterized a variety of aptamers for eight amino acids: arginine, histidine,isoleucine, phenylalanine, tryptophan, tyrosine, leucine, and glutamine (Fig. 3).

Yarus *et al.* argued that, for six of the eight amino acids they reported (all but leucine and glutamine), "coding triplets [codons]…were unexpectedly frequent in cognate RNA-amino acid binding sites" [11]. Given their null hypothesis, namely "that cognate coding triplets are equally frequent everywhere [in any RNA strand], inside and outside RNA binding sites" [11], they argued that "there is no doubt that cognate coding triplets are disproportionately present in the simplest RNA-binding sites for amino acids" [11]. Thus, they conclude, "there

was likely a stereochemical era during the evolution of the genetic code, relying on chemical interactions between amino acids and the tertiary structures of RNA binding sites" [11].

It is this conclusion that some commentators cite as refuting the "very heart" of intelligent design [13] and as "a serious flaw" [14] in design arguments. *Contra* design, these commentators assert, chemistry alone constructs biological information, because undirected stereochemistry can build the first stages of the genetic code.

But is this true? To see why it is not, we shall examine the problems with Yarus *et al.*'s DRT model, in order of increasing severity.[1]

Statistical significance of the DRT model

Yarus *et al.* argue that code-relevant triplets (for cognate amino acids) occur far more than expected, assuming an equal frequency distribution, in the amino acid binding sites of the RNAs that they isolated. This is a statistical significance argument, and like all such, depends critically on background assumptions and the available data. Significance can evaporate, given a different null hypothesis and/or a larger data set. SELEX methods present just such a problem.

Think about the difficulty this way. Suppose one works on a commercial fishing boat, which uses a trawl net. Every day the trawl brings in all kinds of items from the sea, only some of which are marketable fish. The remainder, the crew tosses back into the ocean.

Now, it would be wrong to describe the daily *total haul* only in terms of the fish the crew *keeps*. In a similar fashion, SELEX methods, starting from large pools of random RNAs, capture many different sequences, and investigators must decide which RNAs to keep and analyze further. Under these circumstances, one must guard carefully against introducing a selection bias. The fish stored in the hold, so to speak, are *not* all the fish the trawl captured. Landweber and Knight describe the potential problem:

> Diverse RNA sequences can perform the same task: in SELEX experiments, dissimilar molecules survive many cycles of harsh selection…Few of these sequences are ever further characterized. Consequently, it is possible to choose post-hoc from the same experiments a set of sequences that either does or does not show any particular desired property. [16]

Ellington *et al.* bring the point home:

> The choice of which aptamers to analyze can also significantly influence the statistical validity of the association…In attempting to establish a connection between aptamers and codons one assumes that the aptamers are the product of random sequence; that is, if there is a bias to be discovered, it should be a bias imposed by nature and not by man. [17]

Have Yarus *et al.* introduced such biases into their statistical analysis? The answer appears to be a troubling 'Yes,' as a careful analysis of their 2009 review article makes clear.

Recall that SELEX methods may capture a diversity of RNA sequences "that perform the same task" [13]. For example, when isolating tryptophan-binding RNAs, Yarus *et al.* found RNA sequences (aptamers) with a conserved region, which they dubbed the CYA Trp motif. But they also found 19 unique sequences that, while binding tryptophan, *lacked* the CYA motif. Yet Yarus *et al.* failed to consider these sequences in their analysis. As they acknowledged:

> Nineteen unique sequences (12% of the pool) that do not contain the conserved elements were not tested… In summary, more than one RNA folding fulfilled the selection requirement. There seem to be several ways to construct an RNA site with affinity for tryptophan. [10]

The SELEX trawl captured several RNA sequences that bind tryptophan. Therefore, to avoid bias, all of these sequences should be analyzed statistically—not simply the motifs that look interesting on the stereochemical hypothesis (i.e., sequences exhibiting a disproportionate representation of code-relevant triplets). Otherwise, the screening criteria may artificially amplify the signal the investigators purport to have found—rather like catching both salmon and mackerel, throwing away the mackerel, and then claiming that the trawl caught only salmon.

Additional evidence that Yarus *et al.* set aside data failing to fit their hypothesis can be found when one examines amino acids that don't appear in the list of eight above [11]. Consider valine, for instance. One might think that Yarus *et al.* had yet to investigate RNA binding affinities for valine, but in fact, they did [7].

So why doesn't valine figure in the 2009 statistical analysis, or in Yarus's book [3] on the subject? They did not find code-relevant triplets in the binding site of the valine aptamer. Here's how they explain their omission of the valine results:

> The prevalent valine site in RNA is an internal loop, 4 over 10 nucleotides. Its derivation did not permit deduction of RNA site nucleotides, so we have not used it below for coding triplet calculations. [11]

This sounds reasonable—except that in 1998, Yarus *et al.* noted that they had *failed to find* cognate triplets in the valine binding site [18]. They explicitly contrasted that negative result with the positive arginine and group I intron results:

[1] Although we will not discuss the problem in any detail, we note here that Yarus *et al.* should be modeling activated amino acids, i.e., those already chemically prepared for peptide coupling reactions. Using free amino acids in SELEX experiments represents yet another biochemically implausible aspect of such methods. Activated amino acids (where the carboxyl group of the amino acid is coupled to the 3'-terminal A nucleotide of its corresponding tRNA, as universally employed in organisms during protein assembly) bind very differently than free amino acids. The point can be illustrated by considering the difficulty of peptide synthesis under artificial (human-directed) conditions: "[F]orming the bonds among the 20 different amino acids a sufficient number of times to synthesize a protein still taxes the ingenuity of synthetic chemists. The main problem is the diversity of functional groups on the amino acid side chains. To prevent the participation of these groups in undesirable reactions during the formation of a desired peptide bond, all such reactive groups must be blocked during the synthesis, and the blocking groups must be removed completely after the synthesis. A large number of blocking groups have been developed, each with advantages and disadvantages…Although the synthesis of many peptides is now routine and performed by automatic instruments, the synthesis of many other peptides requires careful consideration of tactics." [15]

> Such functional coding triplets were not found in the selected valine site (Majerfeld & Yarus, 1994), but are frequent among *in vitro*-selected arginine-binding RNAs (Yarus, 1998) and have been found in a natural binding site, the group I intron (Yarus, 1998)…This supports a stereochemical basis for the genetic code….[18]

Further, in the same 1998 paper, Yarus *et al.* describe the nucleotide composition of the valine binding site. "Similarities between the valine and isoleucine sites are easily found," they write. "Both contain conserved strings of G's with apposed U's" [18].

Thus, the claim of Yarus *et al.* 2009 [11], that the valine RNA aptamer results were too poorly characterized to allow their inclusion in the statistical analysis, appears to be contradicted by their earlier publications. Omitting the valine data biases the 2009 analysis in favor of the stereochemical hypothesis.

This biasing underscores another issue. Yarus *et al.* [11] used the wrong null hypothesis to demonstrate codon specificity. They tested for a higher concentration of cognate codons in the amino-acid binding sites of their aptamers, as opposed to the non-amino acid binding nucleotides of the aptamers. But this would be the correct null hypothesis only if Yarus *et al.* had examined *all* relevant RNA sequences (aptamers).

The correct null hypothesis asks whether *non-cognate* triplets are found as often as cognate triplets in the binding sites of *all* aptamers for a given amino acid. However, because Yarus *et al.* evince little curiosity about those unique aptamers that bound amino acids, yet lacked conserved sequence motifs, it is impossible to use the correct null hypothesis. The other sequences have already been tossed back into the ocean. The null hypothesis Yarus *et al.* [11] actually employed, therefore, asks only about the frequency of cognate triplets in the binding sites of the aptamers that they selected for analysis – which looks exactly like the sort of illegitimate statistical bias Ellington *et al.* described as "imposed…by man" [17].

Turning defects into virtues

To establish some chemical affinity Yarus *et al.* must not only show that specific amino acids bind to RNA aptamers, but that amino acids are binding to the RNA where their cognate codons are present or disproportionately concentrated. Yet their own results show more failure than success in establishing a concentration of relevant triplets in aptamer binding sites. Indeed, Yarus *et al.*'s experiments show no chemical affinity between specific triplets and their cognate amino acids in 79% of the RNA molecules they studied. As they note, "…a majority of these experiments (e.g., 79% of specific triplets) have negative outcomes." [11]

They continue:

> Our eight amino acids [see list, above] potentially employ 24 codons and 24 complementary anticodons… Of the possible individual triplets, only 3 of the 24 codons and 7 of the 24 anticodons are significantly found within amino acid binding sites. Thus use of triplets is sparse, as one might perhaps expect…[11]

"As one might perhaps expect?" Defects, remarkably, become virtues in the DRT model:

> These [negative outcomes] can be taken as negative controls, suggesting that these procedures are not strongly biased to find triplets in some profoundly cryptic way. [11]

Or, perhaps, the 'affinities' seen in the aptamer experiments are little more than coincidental patterns, no more causally significant than animal shapes seen in clouds or beach sand. Koonin and Novozhilov [2] note that the 'signal' of the Yarus *et al.* aptamer-amino acid binding results is weak, and also note that Yarus *et al.*'s own results expose another problem for the stereochemical hypothesis. Both codons *and* anticodons show up in many aptamer binding sites, yet there is no plausible mechanism that would allow both the codon and the anticodon to play a role in translation at the same time. As Koonin and Novozhilov note,

> …the affinities are rather weak, so that even the conclusions on their reality hinge on the adopted statistical models. Even more disturbing, for different amino acids, the aptamers show enrichment for either codon or anticodon sequence or even for both, a lack of coherence that is hard to reconcile with these interactions being the physical basis of the code. [2]

Yarus *et al.* are undaunted, however, because they say that stereochemistry can expect a helping hand from other hypotheses, such as coevolution [19] or adaptive optimization [20], to supply the missing triplets. Here, Koonin and Novozhilov shrug at the narrative prowess exhibited:

> Such a composite theory is extremely flexible and consequently can "explain" just about anything by optimizing the relative contributions of different processes to fit the structure of the standard code. Of course, the falsifiability or, more generally, testability of such an overadjusted scenario become issues of concern. [2]

But when we take a look at other shortcomings of the DRT model, the "dubious relevance" of the model for code evolution becomes even more problematic.

The DRT model and modern ribosomal structure

In modern ribosomes, the peptidyl transferase center (PTC), where the peptide bond forms between amino acids carried by tRNAs, is remarkable for its precise three-dimensional geometry. This space, universally conserved in all ribosomes (including mitochondrial ribosomes), enables what Nobel Laureate Ada Yonath calls "positional catalysis," namely, the exact positioning and movement of the amino acid-bearing CCA stems of adjacent tRNAs to enable peptide bond formation at the heart of the ribosome molecular machine [21]. (See Fig. 1 of [21] for an illustration of the ribosome's structure and functional sites.)

No such precision exists in the DRT model. Thus, even if an ensemble of RNA aptamers aligned in close proximity to one

another, and even if they did so in a way that would in theory specify an amino acid sequence with biological relevance (a dubious proposition, see below), no evidence shows that amino acids thus carried by the RNA aptamers would form *peptide bonds*, especially in any realistic prebiotic setting.

In extant cells, the tRNAs that hold amino acids in place for peptide bond formation do so using *covalent* bonds. These strong chemical attachments enable the tRNA to present the amino acid at a distance from the main body of the tRNA molecule, to prevent any steric hindrance to peptide bond formation. The DRT model RNA aptamers, however, bind amino acids using weaker non-covalent associations. As a result, the RNA aptamers *have* to make more extensive contact with their amino-acid ligands.

This raises the possibility that the RNA aptamers will either partially, or completely, envelope the amino acids to which they are bound, or that they will otherwise introduce steric hindrance to peptide bond formation.[2] Recognizing this problem, Yarus *et. al* have carefully engineered their aptamers to ensure that they attach to the side groups of their corresponding amino acids, rather than only to the α-amino and α-carboxyl groups, where peptide bonds form. This engineering clearly represents intelligent design, and thus does not simulate an undirected stereochemical origin of the genetic code, but rather its opposite.

These facts should give pause to the reader. If a "sloppier" and less precise system of stereochemical templating would actually work to build proteins, where is the evidence for that, and why did the modern system ever develop? All protein assembly machines for which we have genuine functioning examples require at least the complexity and precision of prokaryotic ribosomes. Hypothetical simpler systems are hypothetical for good reason: they have not been shown to work.

Figure 4, taken from [11], represents the first stage in the DRT model. From the figure, it appears that the RNA template aligns the amino acids (circle, square, and triangle, with their leaving groups represented as small black ovals) neatly for peptide bond formation. But this arrangement does not fit with the actual architecture of modern ribosomes.[3] As Andrew Ellington and colleagues at the University of Texas-Austin note, the entire schema of Yarus *et al.* would need to be stood on its head, moving amino acids *away* from the RNAs binding them, to make the functional transition from a hypothetical direct RNA template to what is observed in ribosomes today. The implausibility of such biochemical 're-engineering,' Ellington *et al.* argue, is self-evident:

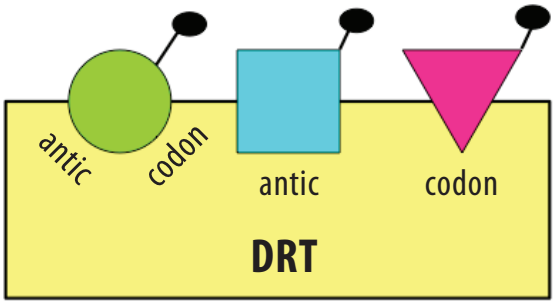

Figure 4. First stage in the DRT model (from Yarus *et al.* [11]). The diagram is *conceptual*, not chemical: circle, square, and triangle represent different amino acids aligned for peptide bond formation, with their leaving groups illustrated by small black ovals; the yellow rectangle is the sequence of RNA aptamers jointly providing a direct RNA template (DRT). "Antic" abbreviates an anticodon sequence. **doi:**10.5048/BIO-C.2011.2.f4

> Thus, if ribosomal RNA is a lineal descendant of these peptide synthetase ribozymes [the Yarus *et al.* aptamers]…then peptide bond formation should still occur adjacent to codons. There is no *a priori*, stereochemical rationale for the separation in space of codons and amino acids, and the large-scale movement of substrates relative to a coevolved active site would be both unnecessary and unprecedented. However, in the modern translation apparatus not only are amino acid substrates and the catalytic core not in direct contact with codons, but amino acids are held rather far away (>70 Å) from codons by a relatively inflexible RNA intermediate, tRNA. It is easier to contradict the [stereochemical affinity] aptamer-codon hypothesis than to invent rationales for how and why tRNA stood up during the course of evolution. [17]

Figure 5 shows what bothers Ellington *et al.*, and should bother anyone who thinks about the massive structural changes

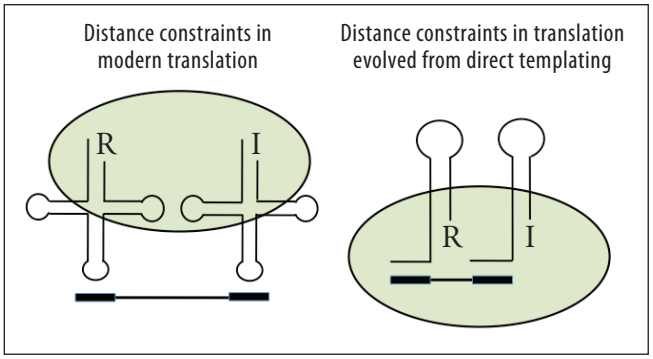

Figure 5. Problem of reversing the orientation of binding templates. To achieve the topology of modern ribosomes (left) would require disrupting the mode of templating in the DRT model (right). R, arginine tRNA; I, isoleucine tRNA. (After Ellington *et al.* 2000 [17].) **doi:**10.5048/BIO-C.2011.2.f5

[2] Indeed, Yarus *et al.* acknowledge that, absent engineering, RNA aptamers will tend to bind to the α-carboxyl and α-amino groups (i.e., the groups where peptide bonds form). In describing how RNA aptamers bind to amino acids, they note that "it is straightforward for RNA to bind an amino acid via its polar features. An important initial implication is that all free amino acids may be RNA bound, because the α-amino and α-carboxyl are always present, supplying good complements to the hydrogen bonding donors and acceptors, for example, at the peripheries of bases, base pairs, and base triples" [11]. In their isolecuine experiments, therefore, Yarus *et al.* took steps to ensure that "glycine was added to both [selection and elution] buffers to discourage enrichment of species with exclusive affinity for the amino group of the ligand" [18]. This intervention is biochemical engineering, with no prebiotic (undirected) analogue. In their phenylalanine experiments, Yarus *et al.* noted that "we selected sidechain-specific RNA binding sites to ensure that the binding target extends beyond interaction with this positive [α-amino] charge" [22] – another instance of engineering, with no prebiotic analogue. Undirected chemistry, which is not looking to construct a genetic code, would not be discriminating.

[3] As Knight and Landweber note, "The main objection to [the DRT model] is that it requires a discontinuity at the point at which adaptors take over from direct templating. Furthermore, it requires that each residue in a peptide be encoded by a large RNA site, but the evolvability of such a system may be limited depending on how specificites are connected in sequence space….There are also potential reading frame difficulties in shifting from many bases per amino acid to only three bases per amino acid" [15].

required to move from a direct RNA template, as postulated by Yarus *et al.*, to modern ribosomes. If aptamers directly positioned amino acids when protein assembly first evolved, then that direct association would need to have been lost, or reversed, to reach what we see today in ribosomes, with codons, anticodons, and amino acids attached to tRNAs at a universal 3' CCA stem. Once again, the "dubious relevance" of the DRT model becomes apparent.

Complex amino acids are an unlikely starting point

If one looks again at the amino acids Yarus *et al.* tested for RNA binding affinities (Fig. 3), one can see that their side chains are relatively large and complex. Smaller amino acids with simpler side chains, such as glycine, serine, or alanine, are missing from the list (and valine failed entirely to correlate with the DRT model). This raises a major difficulty for the DRT model.

Recall that the biosynthetic pathway to a large and relatively complicated amino acid, such as tryptophan, is anything but simple. Surely it would be more plausible, under the DRT model, to try to find aptamers binding the simpler amino acids first? If stereochemical affinities caused the first genetic code, then we should expect to see those affinities in the easiest-to-synthesize amino acids, not in amino acids requiring elaborate, functionally-integrated biosynthetic pathways.

Because those biosynthetic pathways involve many enzymes, extant cells would require a pre-existing translation system in order to make them. Since attempts to explain the origin of the genetic code are also attempts to explain the origin of the translation system (indeed, there can be no translation without a code), Yarus *et al.*'s findings raise an acute chicken and egg problem. Which came first, the aptamer-amino acid affinities that Yarus *et al.* propose as the basis of the code and translation system, or the translation system that would have been necessary to produce those amino acids (and, thus aptamer-amino acid affinities) in the first place?

Although Yarus *et al.* did not find evidence of aptamer-amino acid affinities for the simplest-to-synthesize amino acids, they profess to find nothing odd about this result. As they explain:

> Finally, it is sometimes thought to be surprising that amino acids like arginine and tryptophan, which have complex biosyntheses, are found to belong to the stereochemical group….However, we do not think these findings raise a new or difficult point. Firstly, replication of RNAs accurately so as to preserve ribonucleotide sequences is among the logical necessities for the evolution of coding and translation. Thus highly organized nucleotide synthesis pathways and energy metabolism must have existed in the environment that saw the development of translation; it seems to add little new complexity to impute a concurrent pathway for synthesis of arginine or tryptophan. Secondly, when little information is available it seems to us particularly important to follow the data, rather than preconceptions for which experimental evidence is absent. [11]

In for an inch, in for a mile, it seems. But it is simply false that "little information is available," as Yarus *et al.* claim. Leslie Orgel once termed the imagining of special prebiotic conditions in order to preserve a favored hypothesis as "pigs can fly" assumptions, precisely because they defy *what is already known* about biochemistry and plausible prebiotic conditions [23]. Koonin and Novozhilov explain that the artificial conditions of many origin-of-life experiments yield a net gain of zero, in terms of genuine understanding, if the results disappear when those conditions are removed:

> …it makes sense to ask: do the analyses described here, focused on the properties and evolution of the code *per se*, have the potential to actually solve the enigma of the code's origin? It appears that such potential is problematic because, out of necessity, to make the problems they address tractable, all studies of the code evolution are performed in formalized and, more or less, artificial settings (be it modeling under a defined set of code transformation or aptamer selection experiments), the relevance of which to the reality of primordial evolution is dubious at best. [2]

Troubled by tryptophan

Recent work from the Yarus lab provides an encouraging counterpoint, however, to the shortcomings we have surveyed above [24], and gives us some hope that Yarus *et al.* may begin to look at the DRT model with more healthy skepticism.

Recall that in their 2005 experiments [10], Yarus *et al.* isolated aptamers binding tryptophan (Trp); the sequences they selected for analysis contained a conserved "CYA motif." But biological functions—such as binding Trp—require sites that are not merely necessary but also *sufficient* to produce a given effect. That a sequence such as the CYA motif is conserved, however, does not show that it is *sufficient* for binding: "conservation finds only invariant sequence elements that are necessary for function, rather than finding a set of sequence elements sufficient for function" [24]. Other sequences and structures, perhaps not conserved, may also be needed.

Thus, Yarus *et al.* wondered if the CYA motif was sufficient to bind Trp, a finding which, if demonstrated, would support the DRT /stereochemical hypothesis. To answer this question, they placed the CYA motif "in a random-sequence background" —i.e., they embedded the motif in longer randomized RNA strands—reasoning that if the sufficient "sequence and structural elements required for function were present…we would predict a large fraction of the resulting sequences to show full activity" [24]. Conversely, if additional elements were needed, the CYA motif alone would fail to bind Trp.

What they found surprised them. "When we tested the sufficiency, as well as the necessity for Trp affinity," they observe, "…the single loop [CYA motif] model failed" [24]. Yarus *et al.* then re-examined their Trp-binding aptamers, and discovered that *non-conserved* elements, "elusive to normal criteria of sequence or structural conservation" [24], were required for function.

The CYA motif was necessary for Trp binding (that is, in the class of CYA motif sequences; as noted above, other aptamers lacking the motif also bound Trp), but was not sufficient.

The consequences of this finding for the DRT model are significant. Statistical estimates of the occurrence of cognate codons in aptamers "depend on an accurate census of an active site" [24]. If non-conserved but functionally necessary sequences and structures in aptamers are overlooked, however, estimates based only on conserved motifs will be incorrect. As Yarus *et al.* note,

> previous estimates of the probability of finding particular types of RNA sites…may be inflated by failing to take into account undetectable, but nonetheless important parts of the active site, such as those revealed here. [24]

These new results underscore the point we made above: data relevant to Yarus *et al.*'s statistical analyses of amino acid-binding aptamers should not have been discarded. We are encouraged that they have begun to rethink the tryptophan analysis.

The DRT model and the sequencing problem

One further aspect of Yarus's work needs clarification and critique. One of the longest-standing and most vexing problems in origin-of-life research is known as the sequencing problem, the problem of explaining the origin of the specifically-arranged sequences of nucleotide bases that provide the genetic information or instructions for building proteins.

Yet, in addition to its other deficiencies it is important to point out that Yarus *et al.* do not solve the sequencing problem, although they do claim to address it indirectly. Instead, Yarus *et al.* attempt to explain the origin of the genetic *code*—or more precisely, one aspect of the translation system, the origin of the associations between certain RNA triplets and their cognate amino acids.

Yarus *et al.* want to demonstrate that particular RNA triplets show chemical affinities to particular amino acids (their cognates in the present-day code). They try to do this by showing that in some RNA strands, individual triplets and their cognate amino acids bind preferentially to each other. They then envision that such affinities initially provided a direct (stereochemical) template for amino acids during protein assembly.

Since Yarus *et al.* think that stereochemical affinities originally caused protein synthesis to occur by direct templating, they also seem to think that solving the problem of the origin of the code would also simultaneously solve the problem of sequencing. But this does not follow. Even if we assume that Yarus *et al.* have succeeded in establishing a stereochemical basis for the associations between RNA triplets and amino acids in the present-day code (which they have not done; see above), they would not have solved the problem of sequencing.

The sequencing problem requires that long RNA strands would need to contain triplets already arranged to bind their cognate amino acids in the precise order necessary to assemble functional proteins. Yarus *et al.* analyzed RNA strands enriched

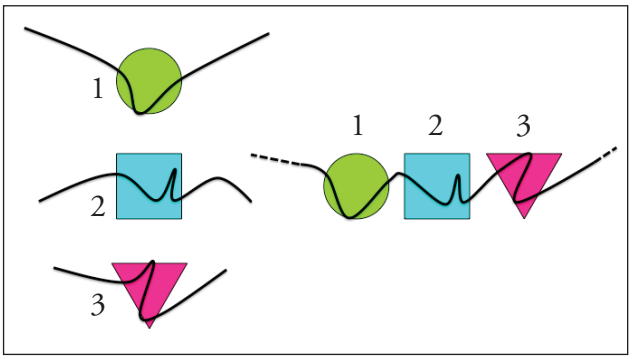

Figure 6. The sequencing problem. In this conceptual (not chemical) diagram, RNA nucleotides (the black strands) which bind amino acids (represented by the green circle, blue square, and magenta triangle) occur in longer aptamers with non-binding bases. To specify protein sequences, which require many different amino acids, code-relevant (i.e., amino acid-binding) nucleotides must be removed from their native aptamers, and re-assembled into new *sequenced* aptamers with correct orientations and molecular distances, to achieve functional sequences of binding sites. **doi:**10.5048/BIO-C.2011.2.f6

in specific code-relevant triplets, and claim to have found that these strands show a chemical affinity with their cognate amino acids. But they did not find RNA strands with a properly sequenced series of triplets, each forming an association with a code-relevant amino acid as the DRT model would require, and arranged in the kind of order required to make functional proteins. To synthesize proteins by direct templating (even assuming the existence of all necessary affinities), the RNA template must have *many properly sequenced* triplets, just as we find in the actual messenger RNA transcripts.

To produce such transcripts, however, would require excising the functional (information-carrying) triplets with code-relevant affinities, from the otherwise non-functional (random) individual RNA strands in which Yarus *et al.* claim to find such triplets, or linking such strands in a way that allows coding (Fig. 6). Functional triplets would need to be concatenated and arranged, to construct something akin to a gene, which could directly template functional proteins. But Yarus *et al.* do not explain how any of this, least of all the specific arrangement of the triplets, would occur. Thus, they fail to solve the sequencing problem.[4] Indeed, even if Yarus *et al.* had succeeded in explaining the origin of triplet-cognate amino acid associations, and even if these associations constituted a fully *functional* code (questionable propositions, as we have indicated), their work would leave unaddressed the crucial sequencing problem. We think it is important to make this limitation in Yarus' work clear, because some scientists, as we discuss below, have claimed that Yarus *et al.* have refuted contemporary versions of the intelligent design hypothesis that do address the sequencing problem.

[4] Of course, Yarus *et al.* might have claimed, more modestly, that the associations they purport to find in their aptamers provide a stereochemical basis for the genetic code. RNA aptamers might originally have functioned, during translation, in the role that transfer RNAs and aminoacyl tRNA synthetases play today. Yet clearly such a system would still need functional genes to provide instructions for building proteins—in which case, the sequencing problem would remain unsolved.

DISCUSSION

The origin of one of the most foundational features of all living organisms—the genetic code—requires careful analysis. Despite the problems with the DRT model described above, some have argued that the model shows that biological information arises directly from chemistry. For this reason, Arthur Hunt and others have claimed that the DRT model of Yarus *et al.* shows the "very heart" of the theory of intelligent design "is wrong" [13]. Citing Yarus *et al.*, Dennis Venema has likewise said that the evidential strength of the DRT model exposes "a serious flaw" in the argument for intelligent design [14], as presented in the recent book written by one of us (Meyer), *Signature in the Cell: DNA and the Evidence for Intelligent Design* (hereafter, *Signature*) [25].

Signature argues that intelligent design provides the best explanation for the origin of the sequence-specific digital information (the genetic text) necessary to produce the first living cell. *Signature* shows, first, that no undirected chemical evolutionary process explains the origin of the information necessary to produce the first life and, second, that intelligent agents, and only intelligent agents, have demonstrated the causal power to produce large amounts of functionally specified digital information (or specified complexity)—at least, starting from purely chemical antecedents. (*Signature*'s argument concerns the efficacy of chemical, not biological, evolutionary processes.) In other words, *Signature* argues that organisms were intelligently designed, because of the presence of specifically-arranged nucleotide bases in DNA and RNA in even the simplest cells; that is, *Signature* addresses *the sequencing problem* (as discussed above) and presents intelligent design as the solution to it.

Nevertheless, as noted above, Yarus and his colleagues neither address, nor solve, that problem. For this reason, they do not refute the case for intelligent design based upon the presence of sequence-specific digital information in DNA and RNA (i.e., the genetic text).

In any case, they do not solve the problem of the origin of the genetic *code* either. Instead, upon analysis, we find:

1. Yarus *et al.*'s methods of selecting amino-acid-binding RNA sequences ignored aptamers that did not contain the sought-after codons or anticodons, biasing their statistical model in favor of the desired results.
2. The DRT model Yarus *et al.* seek to prove is fundamentally flawed, since it would demonstrate a chemical attraction between amino acids and codons that does not form the basis of the modern code.
3. The reported results exhibited a 79% failure rate, casting doubt on the legitimacy of the "correct" results.
4. Having persuaded themselves that they explained far more than they actually had, Yarus *et al.* then simply assumed a naturalistic chemical origin for various complex biochemicals, even though there is no evidence at present for such abiotic pathways.

To be sure, noting the inadequacies of Yarus *et al.*'s DRT model does not constitute a case *for* intelligent design. But our review does show that Yarus *et al.* have neither refuted the specific arguments for ID developed in *Signature in the Cell*, nor foreclosed the possibility that intelligent design might after all provide the best explanation for the origin of the genetic code as well.

One could argue, of course, that the inability to make progress on the longstanding problem of the origin of the code merely indicates that more work is needed. One might argue that given more time, models based solely on the interplay of undirected chance and necessity [26] will eventually solve this problem, and thus that chance and necessity should be left standing as the sole framework for inquiry.

Given, however, the repeated failures to account for the origin of the code within this essentially materialistic framework, it may well be time to consider other approaches.

We see three reasons for so doing:

1. Persistent lack of progress on a scientific problem is exactly what one should expect when a causal puzzle has been fundamentally misconceived, or when the toolkit employed in causal explanation is too limited.
2. Our knowledge of cause and effect, long understood to be the basis of all scientific inference and explanation, affirms that true codes—and the semantic relationships they embody—always arise from intelligent causes. The methodological principle here finds its roots in Isaac Newton's First and Second Rules of Reasoning in Philosophy [27]; in particular, his second, which states, "…*to the same natural effects we must, as far as possible, assign the same causes.*"

   If the genetic code as an effect gives evidence of irreducible semantic or functional mappings—i.e., if what we see operating in cells is not *like* a code, but genuinely *is* a code—then we should seek its explanation in the only cause "true and sufficient" to such effects: intelligence. Moreover, we should expect that hypotheses employing causes other than intelligence will collapse under the weight of unexplained data. Anything that does not actually *cause* **x**, cannot *explain* **x**.
3. To the extent that Yarus *et al.* succeed in establishing any biologically relevant associations between base triplets and cognate amino acids—any correspondences reminiscent of the actual code—they did so as a result of their own intelligent intervention.

Yarus himself, of course, thinks that his work establishes that "it is not credulous" [3] to think that natural processes may explain the origin of the genetic code. But it *would* be credulous to see a natural process at work in what is, in fact, an intelligently-directed or manipulated experiment. As Robert Shapiro has forcefully argued, origin-of-life experiments succeed to the degree that cheating (intelligent intervention) is implicated. "In every case," he notes, "the result was due to the flagrant inter-

ference of the investigator in biasing the results to attain the results that he wanted"[5]. Moreover, it would be credulous to sift the results of such experiments with a target in mind, throwing away data that do not fit one's preferred scenario. To the degree that Yarus *et al.* have done this, they simulate, not the power of chemical affinity, but the need for intelligent design, to generate the semantic associations that constitute actual codes.

[5] Brockman J, ed (2007) Life, What A Concept! at http://www.edge.org/documents/life/shapiro_index.html.

Acknowledgements

The authors gratefully acknowledge helpful discussions with D. Axe, A. Gauger, C. Luskin, J. Wells, and R. Sternberg, and the illuminating assistance of two anonymous referees. Any errors are the authors' responsibility.

1. Judson HF (1996) The Eighth Day of Creation: Makers of the Revolution in Biology. Cold Spring Harbor Press (Cold Spring Harbor, NY).
2. Koonin E, Novozhilov A (2009) Origin and evolution of the genetic code: The universal enigma. IUBMB Life 61: 99-111. **doi:**10.1002/iub.146
3. Yarus M (2010) Life From an RNA World: The Ancestor Within. Harvard University Press (Cambridge, MA).
4. Yarus M (1988) A specific amino acid binding site composed of RNA. Science 240:1751-1758. **doi:**10.1126/science.3381099
5. Yarus M, Christian EL (1989) Genetic code origins. Nature 342:349-350. **doi:**10.1038/342349b0
6. Yarus, M (1991) An RNA-amino acid complex and the origin of the genetic code. New Biol 3: 183-189.
7. Majerfeld I, Yarus M (1994) An RNA pocket for an aliphatic hydrophobe. Nat Struct Biol 1:287-292. **doi:**10.1038/nsb0594-287
8. Yarus M (2001) On translation by RNAs alone. Cold Spring Harb Sym 66:207-215. **doi:**10.1101.sqb.2001.66.207
9. Yarus M, Caporaso JG, Knight R (2005) Origins of the genetic code: The escaped triplet theory. Annu Rev of Biochem 74:179-98. **doi:**10.1146/annurev.biochem.74.082803.133119
10. Majerfeld I, Yarus M (2005) A diminutive and specific RNA binding site for L-tryptophan. Nucleic Acids Res 33:5482-5493. **doi:**10.1093/nar/gki861
11. Yarus M, Widmann J, Knight R (2009) RNA-amino acid binding: A stereochemical era for the genetic code. J Mol Evol 69:406-429. **doi:**10.1007/s00239-009-9270-1
12. Yockey H (1992) Information Theory and Molecular Biology. Cambridge University Press (Cambridge).
13. Hunt A (2010) Signature in the cell? http://aghunt.wordpress.com/2010/01/03/signature-in-the-cell/
14. Venema D (2010) Seeking a signature: Essay review of *Signature in the Cell*. Perspectives on Science and Christian Faith 62:276-83. http://www.asa3.org/ASA/PSCF/2010/PSCF12-10Venema.pdf
15. Creighton T (1993) Proteins: Structures and Molecular Properties. W.H. Freeman (New York).
16. Landweber L, Knight R (2000) Guilt by association: The arginine case revisited. RNA 6:499-510. **doi:**10.1017/S1355838200000145
17. Ellington AD, Khrapov M, Shaw AC (2000) The scene of a frozen accident. RNA 6:485-498. **doi:**10.1017/S1355838200000224
18. Majerfeld I, Yarus M (1998) Isoleucine: RNA sites with associated coding sequences. RNA 4:471-478.
19. Wong JT-F (1975) A co-evolution theory of the genetic code. Proc Natl Acad Sci 72:1909-1912. **doi:**10.1073/pnas.72.5.1909
20. Freeland SJ, Hurst LD (1998) The genetic code is one in a million. J Mol Evol 47:238-248. **doi:**10.1007/PL00006381
21. Yonath, A (2009) Ribosome: An ancient cellular nano-machine for genetic code translation. In Puglisi JD, ed., Biophysics and the Challenges of Emerging Threats. Springer-Verlag (New York). pp. 121-155. **doi:**10.1007/978-90-481-2368-1_8
22. Illangasekare M, Yarus M (2002) Phenylalanine-binding RNAs and genetic code evolution. J Mol Evol 54:298-311. **doi:**10.1007/s00239-001-0045-6
23. Orgel L (2008) The implausibility of metabolic cycles on the prebiotic Earth. PL oS Biol 6:e18. **doi:**10.1371/journal.pbio.0060018
24. Majerfeld I, Chocholousova J, Malaiya V, Widmann J, McDonald D, Reeder J, Iyer M, Illangasekare M, Yarus M, Knight R (2010) Nucleotides that are essential but not conserved; a sufficient L-tryptophan site in RNA. RNA 16:1915-24. **doi:**10.1261/rna.2220210
25. Meyer S (2009) Signature in the Cell: DNA and the Evidence for Intelligent Design. Harper One (New York).
26. Monod J (1971) Chance and Necessity. Vintage (New York).
27. Cajori F (1962) Sir Isaac Newton's Mathematical Principles of Natural Philosophy, trans. Andrew Motte (1729). University of California Press (Berkeley).