The evolution of the genetic code: impasses and challenges

Ádám Kun[1,2,3] & Ádám Radványi[4]

[1] Parmenides Center for the Conceptual Foundations of Science, Munich/Pullach, Germany.
[2] MTA-ELTE Theoretical Biology and Evolutionary Ecology Research Group, Budapest, Hungary.
[3] Evolutionary Systems Research Group, Centre for Ecological Research, Tihany, Hungary
[4] Department of Plant Systematics, Ecology and Theoretical Biology, Institute of Biology, Eötvös University, Pázmány Péter sétány 1/C, 1117 Budapest, Hungary

Abstract

The origin of the genetic code and translation is a "notoriously difficult problem". In this survey we present a list of questions that a full theory of the genetic code needs to answer. We assess the leading hypotheses according to these criteria. The stereochemical, the coding coenzyme handle, the coevolution, the four-column theory, the error minimization and the frozen accident hypotheses are discussed. The integration of these hypotheses can account for the origin of the genetic code. But experiments are badly needed. Thus we suggest a host of experiments that could (in)validate some of the models. We focus especially on the coding coenzyme handle hypothesis (CCH). The CCH suggests that amino acids attached to RNA handles enhanced catalytic activities of ribozymes. Alternatively, amino acids without handles or with a handle consisting of a single adenine, like in contemporary coenzymes could have been employed. All three scenarios can be tested in *in vitro* compartmentalized systems.

Keywords

Origin of Life; genetic code; RNA world; ribozyme; coding coenzyme handle;

Introduction

Modern cells store information in DNA and have peptide enzymes to carry out the metabolism for the cell. The information stored in DNA sequences are translated to protein sequences via the process known as translation. During translation a messenger RNA (mRNA) is transcribed from the DNA. The mRNA attaches to the ribosome (rRNA), an RNA-peptide complex that catalyses the RNA dependent polymerization of amino acids. The amino acids are carried to the ribosome by transfer RNAs (tRNA). Thus between DNA and peptides we find a host of RNAs. This fact has already sparked the mind of Francis Crick to propose an RNA world (Crick, 1968), in which RNA acts both as information storing molecule and as enzymes. Naturally occurring RNA enzymes were found in the early '80s (Guerrier-Takada et al., 1983; Kruger et al., 1982) giving more credit to the idea of an RNA world. The fact that the ribosome is an RNA enzyme, in which the peptides act as scaffolds and regulators, suggests that is was evolved during the RNA world era of the origin of life.

The RNA world hypothesis was gradually elaborated and developed. While the puzzle of this stage of the origin of life needs some more pieces (Kun et al., 2015), the picture is getting clear, and now the RNA world hypothesis is the most accepted among the scholars.

The basic idea of the RNA world was established in the '60s and '70s (Lazcano, 2010), the term itself originates from the '80s (Gilbert, 1986). Still, different researchers can mean different things when they refer to the RNA world. For some, it only encompasses the pre-cellular stage, i.e. the formation and possible evolution of the macromolecules (RNA and lipids) and surface metabolism. Thus for them RNA world is something not really alive and very primitive. For others, including us, the RNA world era includes the appearance of the first cell. Moreover, ribocells could have achieved quite complex metabolism (Kun et al., 2015), which is required for translation. Translation is not the sole possibility of employing amino acids or even peptides by the primordial metabolism. Some of the scenarios presented below propose that amino acids and/or peptides were employed from an early stage of the RNA world. That very well could be the case. As long as RNA acts as the main information storage and most of the catalysis is done by ribozymes, we are still in the RNA world. The RNA world era has witnessed some of the most fundamental innovations for Life, the first living cell among them.

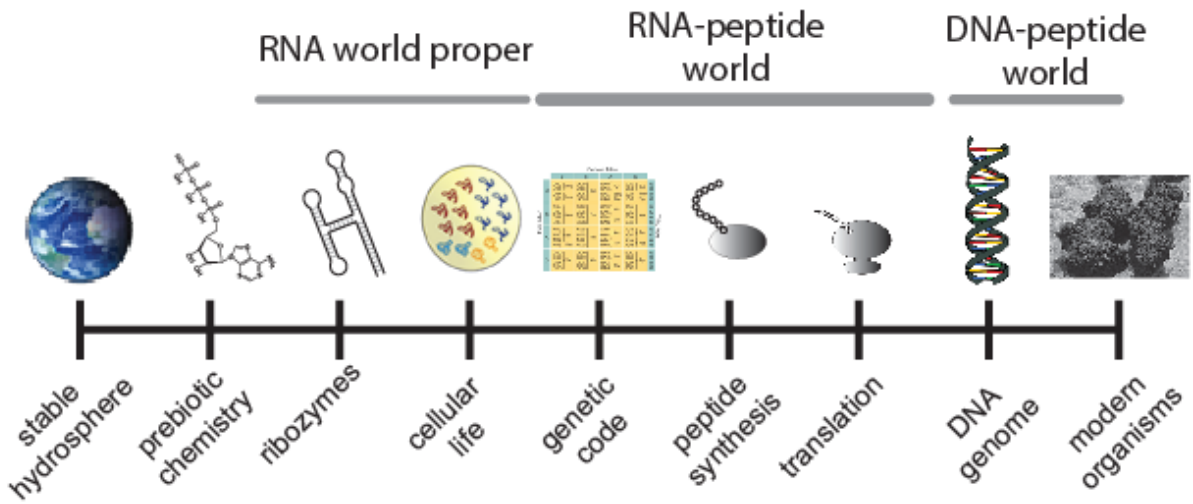

Figure 1. A conceptual figure of the stages of the origin of life, with a focus on the RNA world and its transition to the DNA-protein world. The order of the appearance of genetic code and peptide synthesis could have also happened in reverse order. We argue for a scenario in which the genetic code established first and amino acid polymerization second.

The RNA world had to give rise to the DNA-protein world. This DNA-protein world is the one we live in, and as such, this is the final phase all origin of life scenarios need to lead to. How can a system in which RNA genes are transcribed to RNA, which then folds to become a functional enzyme, transit to one in which the transcribed RNA is used to code for a protein? A functional ribozyme and an RNA having an orderly sequence of triplets coding for amino acids are two very different beasts. Furthermore, translation requires some hundred genes (Gil et al., 2004), a good portion of a minimal gene set of bacteria. One needs to savour the irony

of this: the discovery of the molecular details of translation led to the formulation of the RNA world hypothesis, in which the evolution of translation is a rather difficult problem. The path out of the RNA world seems to be very hard. Not only a host of new enzymes need to evolve, mRNAs (coding sequences) have to appear. It is no wonder that the evolution of translation was labelled as a "notoriously difficult problem" (Crick et al., 1976). 40 years later it is still an unsolved cluster of questions.

Difficult problems should be chopped up to more palatable chunks. Szathmáry proposed more than two decades ago that the evolution of translation should be separated into two problems (Szathmáry, 1990, 1993): (1) the origin of the genetic code, i.e. the map between amino acids and triplets of nucleotides; and (2) the origin of translation, i.e. polymerization of amino acids based on information coded in nucleic acids. Now we think that the second part could be further divided into (2a) the evolution of polymerization of amino acids (Krupkin et al., 2011; Noller, 2004); and (2b) the evolution of mRNA and that of translation. In this review, we focus on the first sub-problem, the evolution of the genetic code. The second part(s) of the problem will be tackled elsewhere.

The evolution of the genetic code is an adaptation, and thus it should be assessed as all other adaptations. We can, for example, adopt the slightly modified definition of Reeve and Sherman: "An adaptation is a *heritable* phenotypic variant that results in the highest fitness among a specified set of variants in a given environment" (Reeve and Sherman, 1993). (In this regard Sober's (Sober, 1984) or West-Eberhard's (West-Eberhard, 1992) definition would work as well.) So the genetic code was a variant that resulted in higher fitness in the environment it had appeared. We have added that the phenotypic variant needs to be heritable; some of the ideas presented for the origin of the genetic code or translation fail at this. We also need to be very specific about how the genetic code increased the fitness of the riboorganism it evolved in. Just assuming that protein enzymes are better than ribozymes in the long run, is not enough. Evolution has no foresight.

Our first aim with this review is to give a list of questions that any full theory of the origin of the genetic code is required to answer. By full theory we mean one that explains the emergence of genetic code and one which is an acceptable adaptation scenario.

- What is the basis or mechanism behind the assignment of amino acids to the codons?
- What is the selective advantage of having amino acids, oligopeptides or polypeptides in the RNA world?
- What is the fitness advantage of the assignment of an amino acid to a codon?
- What is the order of inclusion of amino acids into the code? I.e. how does it evolve?
- Why this particular set of 20 amino acids is in the code?

Amino acids could have been present on primordial Earth (Bada, 2013), some even in great quantities. In order to be incorporated into primordial metabolism or a ribocell, amino acids had to play a role. This role could have been the enhancement of reactions. This is the most widely made assumption as polypeptides make up contemporary enzymes. Other functions could also be envisioned, like scaffolding (Noller, 2004), membrane transport (Morris, 2002), UV protection (Doig, 2017), energy storage (de Vladar, 2012), etc. The first functional amino

acid containing entities might not have been individual amino acids, but dipeptides or even sort oligopeptides. Ikehara (Ikehara, 2016), for example, proposed that oligomers of the VADG amino acids were the first to appear. The longer peptides first scenarios are problematic as while one can select on random sequences, without heredity there could be no evolution (Zachar and Szathmáry, 2010), i.e. retention of those sequences for further generations.

A genetic code is an assignment of triplets to amino acids. How was a given triplet chosen to code for a certain amino acid? The answer can range from random, through based on physicochemistry or dictated by biochemistry. Furthermore, why assign amino acids to codons? Just because it will come handy once there is translation, it will not be selected. The binding of amino acids to codon sequences or to small handles / adaptor bearing the anticodon triplet had to possess some fitness advantage.

Finally, a theory of the origin of the genetic code should also deal with how the code evolved and populated by novel amino acids. The more amino acids present in the genetic code, the more diverse the function of single amino acids, oligopeptides or proteins. Why this particular set of 20 amino acids is coded in the genetic code? There are numerous other α-amino acids even in our metabolism, like ornithine, which are not coded. And why only 20 (22 if we also take selenocysteine and pyrrolysine into account) amino acids?

The second aim of the review is to propose empirical test for the hypotheses. There is a scarcity of experimental results for this important problem. The techniques needed to test some of the hypotheses are available and should be done in the near future.

This essay does not aim to cover all the literature. We focus on our two goals and the advances brought about in the last few years. Excellent reviews cover the prior literature (Barbieri, 2015; Koonin, 2017; Wong et al., 2016; Yarus, 2017).

The Stereochemical Hypothesis

The stereochemical hypothesis states that there is a stereochemical basis for the assignment of a given codon to an amino acid. As the genetic code can change (even if slightly), even if such physiochemically determined assignment exist, it can be overridden by the molecular machinery. This is important, as the genetic code is a true code and as such it needs to be arbitrary (Barbieri, 2015). But at certain stages of the evolution of the genetic code, physicochemical characteristic could have played a role.

This hypothesis is also quite old, past its 50th birthday. It was shown in 1966 that amino acids have different mobility on paper chromatography in the presence of pyridine (Woese et al., 1966a). Woese and co-workers then proposed that not only pyridine but nucleotides could also bind amino acids differently. With regard to the genetic code, specific binding of an amino acid to its cognate codon or anticodon could explain some of the assignment of codons in the genetic code (Woese et al., 1966b).

There are two line of research that shows that at least for some of the amino acid-(anti)codon pairs there is a stereochemical binding. One focuses on binding by triplet or minihelixes to the amino acids, the other focuses on the binding site of evolved aptamers.

The main line of research to show stereochemical affinity of the cognate anticodon (or the codon) to the amino acid is based on aptamers and the study of RNA-amino acid binding.

Initial attempts were motivated by the *Tetrahymena* self-splicing group I intron, which binds arginine (Yarus, 1988). After 30 years of work, Yarus and colleagues created a well-founded framework for the stereochemical hypothesis (Yarus, 2017; Yarus et al., 2009). He and his co-workers have evolved aptamers to bind amino acids. The binding site contains (enriched in) the codon and/or the anticodon for the cognate amino acid. This was shown for arginine (Connell et al., 1993; Janas et al., 2010); histidine (Majerfeld et al., 2005; Turk-MacLeod et al., 2012); tryptophan (Majerfeld and Yarus, 2005); isoleucine (Legiewicz and Yarus, 2005; Lozupone et al., 2003; Majerfeld and Yarus, 1998) (some of the aptamers can also discriminate against valine and norleucine); phenylalanine (Illangasekare and Yarus, 2002); and tyrosine (Mannironi et al., 2000). On the other hand, no such enrichment were observed for leucine, valine and glutamine. Furthermore, it is not yet settled whether the amino acids should attach to their codons or to their anticodons (Szathmáry, 1999).

Another line of research focuses on triplets and their affinity to their cognate amino acids. There is a discrimination of hairpins with anticodon loops toward their cognate amino acids. Gly was selectively attached to a hairpin bearing its own anticodon as opposed to Phe or Trp anticodons; similarly Ala was selectively attached to its hairpin as opposed to Ser of Phe anticodon bearing hairpins (Shimizu, 1995). Furthermore, cysteine, arginine, methionine and valine preferentially attach to the CGUA and AUGC sequence containing their codons. The other amino acids, except for phenylalanine, have a lower affinity to these sequences (Root-Bernstein, 2010).

Thus the stereochemical hypotheses can account for the codon assignment of Arg, His, Ile, Phe, Tyr, Trp, Gly, Ser, Ala, Cys, Met. No such interaction was found for Gln and Val, and no research was conducted for Lys, Asp, Glu, Thr, Asn and Pro. This hypothesis is very clear on one of our questions: the assignment of codons to amino acids. These amino acids or a subset of them should have been the first to enter the code. The theory does not say anything about the fitness advantage of amino acids, or of the assignment. Implicitly it states that some of the 20 amino acids entered the code because of the stereochemical binding. It does not require that all amino acids behave this way, and for at least two coded amino acids (Gln, Val) no such affinity was detected. It leaves the later part of the evolution of the genetic code to other hypotheses (Yarus et al., 2005).

For a complete picture, all amino acids need to be tested. Actually, what needs to be shown is that the anticodon (or the codon) has a significantly higher affinity to the amino acid or a significantly higher probability of appearance in the binding site for the amino acid than for all other amino acids. Root-Bernstein has shown (Root-Bernstein, 2010) that phenylalanine has a high affinity to the codon of Cys, Arg, Met and Val. Thus stereochemistry cannot distinguish phenylalanine. Only anticodons / codons that can discriminate among the amino acids are good candidates. In order to be a truly good experiment in favour of the stereochemical theory, the experiment of Root-Bernstein should be redone with triplets and for all possible codon and amino acid pairings.

The Coding Coenzyme Handle Hypothesis

The Coding Coenzyme Handle Hypothesis (CCH) (Szathmáry, 1990, 1993, 1996, 1999) tries to answer the question that the previous (and pretty much all other) scenario leaves open: what was the adaptive advantage of the genetic code? Instead of full-blown polypeptides, individual amino acids could have also acted as catalysts, or at least lend their diverse side-chains to the catalytic core of ribozymes. Thus, the amino acids act as coenzymes to the ribozymes. The genetic code enters the picture by attaching the amino acid to a RNA handle. This handle (a proto-tRNA) then can attach to ribozymes via its triplet bearing endloop. Thus the triplet and the amino acid are linked, and ribozymes can employ the right amino acid by selectively binding through the triplet.

There is experimental evidence that shows catalytic help of ribozymes by amino acids. Unfortunately, there is only one such evidence. Roth and Breaker (Roth and Breaker, 1998) has isolated a deoxyribozyme that used the amino acid histidine as a cofactor. L-histidine was added to the reaction mixture, and some of the isolated cleaving deoxyribozymes were not functional without its presence, thus demonstrating that histidine was used as a cofactor. The specificity for employing histidine by the evolved deoxyribozyme was quite high, histidine analogues or histidine containing dipeptides failed to show catalysis, except for L-histidine methyl ester. Thus a single amino acid positioned in the right way can catalyse a reaction.

Modern enzyme centres are 3-dimensional “pockets” in which certain residues are positioned in the right way to form the active site catalysing the reaction. Actually, not many amino acids constitute an active site (Porter et al., 2004). In a simplified concept of enzymes, it is irrelevant whether the role of skeleton is filled by protein or ribonucleic acid, and the reaction itself is only catalysed by the active site. If the substrate is already bound by an aptameric RNA site, only the functional group and its proper orientation is required for facilitation.

We have collected active sites from the Catalytic Site Atlas (Furnham et al., 2014) that only have a single amino acid in them. Then their functionality was determined by linking the entries with the PDB (Berman et al., 2003). There were 123 functional peptides in the database with a single amino acid at their active sites. These peptides span all main enzymatic classes (Fig. 2a). Thus a high metabolic diversity could be maintained with the proper positioning of a single amino acids. The amino acids thus employed (Fig.2b) are mainly the ones that are frequent in active sites (Kun et al., 2007), but their order of abundance is different. Histidine is the most abundant amino acids in active sites, but glutamic acid and aspartic acid more often act alone. Aspartic acid will be important later on. But we are not interested in peptide enzymes, but ribozymes with amino acid cofactors. We have seen that by the positioning of a single amino acid catalysis can be achieved. Thus we can assume that if the binding of the substrate, and positioning is done by RNA instead of a polypeptide, an amino acid can still catalyse reactions as a cofactor.

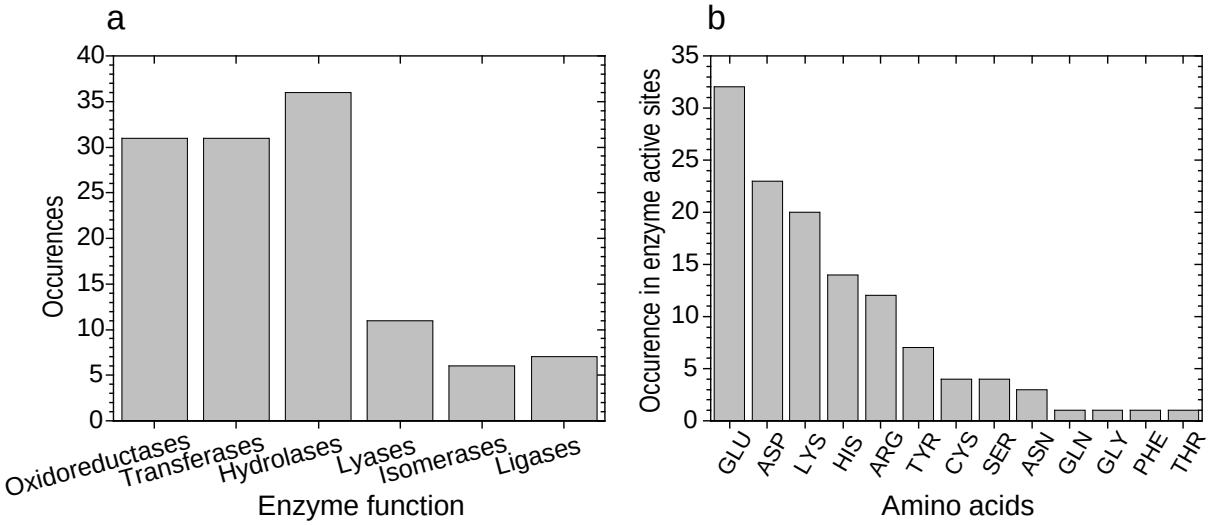

Figure 2. Enzyme centres with only one amino acid. (a) The distribution of enzymes with one amino acid at their active centrums distributed among the main E.C. classes. (b) The occurrence of single amino acids in the active centrums of enzymes.

However, amino acids are not magic substrates that can enhance catalytic activity of any ribozyme by merely attaching them to the ribozymes. Müller and coworkers (Yao et al., 2011) attached an arginine to small oligonucleotides, and by this handle they probed if arginine could enhance the activity of the RNA polymerase ribozyme (Johnston et al., 2001). They reported a negative result: „arginine did not improve polymerization when placed at ten different positions on the polymerase ribozyme" (Yao et al., 2011). This might be seen as a blow to the CCH, but the authors rightfully argue that "the current polymerase ribozymes may not benefit from the conjugates because the ribozymes were optimized in the absence of these conjugates" (Yao et al., 2011). Without the trouble of coevolving the amino acid coenzymes and the ribozymes, we cannot expect any enhancement of catalytic activity.

The experimental result of histidine as ribozyme cofactor (Roth and Breaker, 1998) is cited as one potentially backing the CCH. On one hand, it proves that a single amino acid can enhance catalytic activity, as the cleavage reaction was also evolved without the help from the histidine. On the other hand, the histidine was not bound to any handle, thus amino acids alone could have acted as cofactors without help from base-pairing. That would help us explain why amino acids were co-opted in the RNA world, but not the evolution of the genetic code.

Cofactors, like NAD, FAD, CoA or SAM, all include an adenine part, which led White to propose that they are relics from an earlier era (White, 1976). Indeed, coenzymes with their adenine part, which latter is not the functional part of the molecule, are one of the testimony for the existence of a prior RNA world. Ribozymes could grab the adenine attached functional molecules through the adenine (Jadhav and Yarus, 2002; Saran et al., 2003). Adenine (AMP) seems to be the only among the four canonical bases that can easily form coenzymes (Jauker et al., 2015), by condensing with a functional molecule. Furthermore, not only nicotinamide mononucleotide or flavin mononucleotide can be condensed with AMP to yield the coenzyme NAD^+ and FAD, respectively, but peptides also condensate readily (Jauker et al., 2015). Thus

adenine-conjugates could have been present in the primordial environment employed by the metabolism of the ribocells. S-adenosyl methionine is a prime example of an AMP-amino acid conjugate, which still serves as a coenzyme.

van der Gulik (Gulik 2015) argues that it was mostly catalytic dipeptides that appeared first, and their coded synthesis was important to produce these catalytic species. The feedback loops he proposes is „GlyGly produces AspGlyAsp and AspGlyAsp-like sequences; these in turn protect and produce RNA; RNA produces GlyGly" (Gulik 2015). He does not suggest any handles to carry the amino acids and/or the oligopeptides, but his theory crucially depends on the production of GlyGly dipeptides. Other catalytic dipeptides are also known: ValAsp and AlaAsp can catalyse the aminoacylation of haripins with anticodon loops (Shimizu, 1995); SerHis and GlyGly can catalyse peptide bond formation (Gorlero et al., 2009); and AlaHis, but His alone not, is able to catalyse peptydiltransfer of Phe, Pro, Lys and Gly if a template was also present (Shimizu, 1996). In the latter experiment, a host of other dipeptides were also assayed, but they showed no activity, thus only some of the dipeptides are catalytic. Such dipeptides can be produced by amino-acylating ribozymes, most probably on a small handle. The shortest known ribozyme (Turk et al., 2010) does just that: trans-aminoacylates a 4nt long substrate, and can catalyse the addition of more amino acids to the first added. Thus implicitly, the catalytic handle is present in this theory as well. Other dipeptide combinations of Val, Ala and Gly are also effective in the stereospecific synthesis of tetroses (Weber and Pizzarello, 2006). There is an interesting non-catalytic adaptive aspect of peptides, which is based on the cellular aspect of riboorganisms. PheLeu dipeptide could bind to vesicle membranes, which thus obtain enhanced affinity for fatty acids and thus promotes vesicle growth. This is clearly an adaptive trait, leading to competitive exclusion (Adamala and Szostak, 2013).

While amino acids and dipeptides can clearly enhance the catalytic repertoire of the RNA world, they do not necessary add to the evolution of the genetic code. If amino acids in themselves can be employed by the ribozymes, or if they are attached to an AMP-handle, then there is no association between a triplet and an amino acid. Such scenarios can explain why and how amino acids were first employed, but does not help to uncover the origin of the genetic code.

Given that an amino acid can help catalyse a reaction, an experimental test of the alternating hypotheses can be put forward. A ribozyme should be evolved to catalyse a reaction in the presence of (a) nothing (this will serve as reference); (b) a single amino acid; (c) the amino acid attached to an AMP; and (d) the amino acid attached to a minihelix, having a corresponding anticodon at its end-loop. The reaction should be one that we know that can be catalysed by a ribozyme. It would be very interesting to show that the inclusion of an amino acid allows the catalytic repertoire of ribozyme to broaden, but then one needs to first demonstrate that said reaction cannot be catalysed by ribozymes. So let us say we choose alcohol dehydrogenase as the target, which can be achieved by a ribozyme (Tsukiji et al., 2003). The *in vitro* selection technique should be *in vitro* compartmentalization (Griffiths and Tawfik, 2006; Miller et al., 2006), which can effectively select for multiple turnover ribozymes (Agresti et al., 2005). We need to stress the requirement for multiple turnover, and ribozymes selected by SELEX can fail in this regard, thus we propose that *in vitro* compartmentalization techniques should be employed. *In vitro* compartmentalized systems

are beginning to be used at an increasing frequency in origin of life research. RNA can be copied and compartments selected based on the activity of the ribozymes within the compartment (e.g. (Matsumura et al., 2016)).

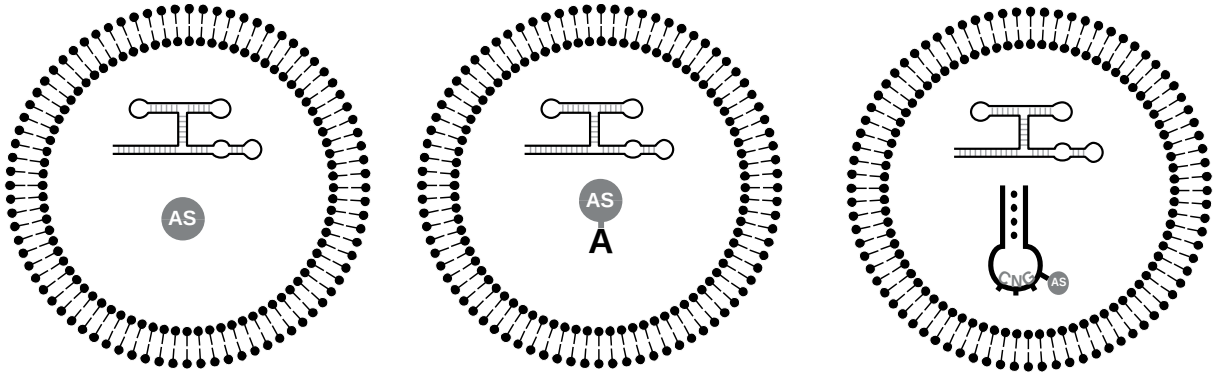

Figure 3. Schematic representation of the suggested experimental setup. RNAs potentially folding to some structure and individual amino acids, AMP bound amino acids, or handle bound amino acids are compartmentalized and selected for catalysis.

We envision the four setups as follows: case (a) and (b) are straightforward. In case (a) the ribozyme is selected without the presence of any amino acids. In essence, it is a repeat of an experiment to select for the chosen catalytic activity to yield a ribozyme. Case (b) is also straightforward as the amino acid needs to be added to the reaction mixture. At first catalytically proficient amino acids should be used, especially Asp. In treatment (c) the amino acid is covalently bound to the nucleotide. As a first choice, the nucleotide should be adenine, and the amino acid attached through a phosphoramidate link. The phosphoramidate link forms as shown by Richert and co-workers (Jauker et al., 2015). Aminoacyl-AMP can also be formed with an anhydride link, as used in the activation of amino acids in contemporary metabolism. As argued below the stronger N-link is preferable. In this scenario, there is a coenzyme (the amino acid adenylate) but it is not a coding coenzyme. The last treatment (d) test the CCH hypothesis. In the original CCH hypothesis the amino acid was attached to a single nucleotide (Szathmáry, 1990) or directly to the anticodon (Szathmáry, 1993). This turned out not to be a satisfactory solution, and later Szathmáry proposed that the handle should be a stem-loop (a minihelix) (Szathmáry, 1996, 1999). Kissing end-loops offer surprisingly strong binding (Bouchard és Legault 2014). Where should the amino acid attach to the handle? One solution would be at the 3’ end of the handle, much like at the end of the acceptor stem of modern tRNA. However, then the amino acid could be positioned too far from the ribozyme. Moreover, the labile anhydride bond employed in modern tRNA might not be stable enough. A N-linked amino acid binds more strongly to the stem. Such bonding can be through a modified nucleotide, one which we find at position 37 very frequently. Fort this reason (explained in more detail in (Kun et al., 2007)) we propose that the amino acid should be attached next to the anticodon triplet in the stem-loop handle.

We believe that in all three setups a ribozyme could be selected. The selected ribozymes need to be ones that actually employ the “coenzymes”, thus if they show reactivity without the amino acid then the amino acid was not employed in the catalysis. If one of the systems is clearly more effective than the others, then we can arrive at a conclusion. If the rate enhancements (basically k_{cat}) are roughly the same in the different setups, then the system

reaching said rate enhancement at an earlier round of selection is the best. The immediate fitness advantage comes from the ease by which the coenzyme is accepted and employed by ribozymes. An evolving system without an enzymatic activity is a metabolic cost to the system. As soon as rate enhancement occurs there is also a benefit to the ribocell.

The Coding Coenzyme Handle hypotheses can answer two of our question: the fitness advantage of amino acids (catalysis) and the assignment to a handle (recognition by ribozymes, employment as coenzymes). We have also proposed (Kun et al., 2007) that catalytically important amino acids, like His, Asp, Glu should have been first to enter the code. While some of the catalytically important amino acids, like Lys and Arg are metabolically complex, this should not be a problem for a metabolically rich ribocell.

The first amino acids in the code: the four-column theory

The proposal that catalytic amino acids were the first to enter the code is in some part unorthodox. In a refined metabolic system, complex amino acids, such as arginine and histidine could have been feasible. These amino acids could have easily benefitted the RNA world as efficient cofactors for ribozymes being catalytically highly promising amino acids. On the other hand, if riboorganisms possessed low metabolic complexity, only the simplest, prebiotically abundant amino acids could have been incorporated. Gly, Asp/Glu, Ala and Val (VADG amino acids for short) are commonly formed in Miller's type reactions (Bada, 2013), which is also supported by samples from hydrothermal vents and meteorite specimens (Longo and Blaber, 2012). In this set only Asp/Glu has high catalytic activity, while the others remain "boring" in this sense (Bartlett et al., 2002; Kun et al., 2007). Yet it is important to note again that enzyme centres containing individual amino acids contained Asp/Glu in the first place, hence high metabolic complexity was not necessary for the first catalytic cofactors.

Glycine might have been the first amino acid to enter the code. Tamura argues that glycine is transferred from glycyl-AMP to tRNA with UCCA 3' end by the action of the latter (Tamura, 2015). Thus the assignment works catalytically. Bernhardt and co-workers (Bernhardt and Patrick, 2014; Bernhardt and Tate, 2008) have also proposed glycine as the first amino acid to enter the code. They argue that tRNA of Gly can be derived from a duplication of hairpins with CCA 3' terminus (Widman et al., 2005). Present day tRNAs have CCA at its acceptor stem, of which a part could be the origin of the NCC anticodon of $tRNA^{Gly}$. But present day tRNA evolved prior to translation, when quite some of the code has already been set.

If we accept that the genetic code was populated incrementally, there was a stage where either some of the triplets did not code for any amino acids or much fewer amino acids were coded. The first four amino acids to enter the code were the VADG amino acids, now occupying GNN codons. How to proceed from this stage forward? Trifonov (Trifonov, 2000; Trifonov, 2004) suggested that the initial VADG amino acids were followed by Pro, Ser, Glu/Leu, Thr, Arg, Asn, Lys, Gln, Ile, Cys, His, Phe, Met, Tyr and Trp, in this order. This ordering of the amino acids is based on physical (e.g. duplex stability of the codon-anticodon interaction), chemical (yield is prebiotic experiments, presence in the Murchison meteorite, number of non-hydrogen atoms, chemical inertness, etc.) and biochemical properties (composition of extant proteins, tRNA characteristic, biosynthesis pathways, etc.) of the

amino acids or their codons/anticodons (Trifonov, 2000). Met and Trp are probably latecomer as they have only one codon (see later), moreover Trp, Cys, Tyr and Phe have increasing frequencies in peptides since LUCA (Brooks et al., 2002), and thus could be late additions to the genetic code. However the background mechanism is hard to grasp in this scenario.

Higgs' four-column theory offers a well-interpreted solution (Higgs, 2009). At first, the genetic code only used the 2nd codon nucleotide for coding, the others were still there because they offered strong binding between codon and anticodon (Eigen and Winkler-Oswatitsch, 1981). If amino acids do not need to code for peptides, then a GNC code in itself could have been the first step in the evolution of the genetic code (Eigen and Winkler-Oswatitsch, 1981; Ikehara et al., 2002; Trifonov, 2004). If there already was coded peptide synthesis then the columns consisting of NUN = Val, NCN = Ala, NAN = Asp/Glu or NGN = Gly codons were employed. Afterwards only amino acids, which have similar physicochemical properties and preserve the error minimalization property of the code (i.e. the cost of translation error with the new amino acids is still low), could have entered the code. In this regard, the four-column theory is very explicit on the fitness advantage of the inclusion of novel amino acids (albeit it depends on the genetic code to be used for translation). It turns out, that Ile and Leu should enter the 1st column (NUN); Thr and to a lesser extent Ser and Pro the second (NCN); and Gln, Asn and Lys should enter the 3rd column (NAN). Among these, Ser, Ile, Leu, Pro and Thr are among the potential first 10 amino acids to enter the code (Higgs and Pudritz, 2009; Trifonov, 2000). Interestingly, Higgs found no amino-acids that should enter the 4th column (NGN). Thus after the stage in which only the middle nucleotide coded for an amino acid, there was a phase when the 1st position also become coding, at least for the first two columns (NUN and NCN). By this stage the first 10 amino acids have entered the code.This scenario coalesces with the proposal of Massey (Massey, 2006) that it was the 2nd, the 1st and then the 3rd position which become coding (the 2-1-3 model).

Population of the genetic code: the coevolution theory

The main theory for the population of the genetic code, the coevolution theory, postulates that since prebiotic synthesis was not a feasible source of all twenty protein amino acids, some of them had to be produced via biosynthesis. The formation of amino acids by biosynthetic pathways guided the development of the genetic code (Di Giulio, 2008; Wong, 2007; Wong, 2005; Wong, 1975). An extension of this theory (Di Giulio, 2008) states that the first amino acids came from biochemical pathways directly originating from sugar degradation. Thus Ala (from pyruvate), Asp (from oxaloacetate) and Ser or Gly (from phosphoglycerate) could have been the first amino acids to be employed. While this set agrees well with the one derived from prebiotic availability, it rests on a metabolically rich ribocell, and requires no prebiotic formation of these amino acids.

The coevolutionary theory, while leaving the first codon assignments to some other mechanism, proposes that later additions to the genetic code should be only one position away from their synthetic precursors. Such a pattern can be found in the genetic code. The four-column theory was unable to decipher the origin of the amino acids assigned to the NGN

codons. Serine can be the precursor Gly, Cys and Trp, whereas Arg can form from Glu (Di Giulio, 2008).

The coevolutionary theory does not say much about the fitness advantage of an enlarged amino acid repertoire. The extension of the genetic alphabet can result in better catalysis. Continuous evolution of a simplified chorismate mutase by expanding its amino acid alphabet from 9 to 20 letters provided an enhanced enzyme variant that has improved protein stability and catalytic activity. This indicates a benefit for fine-tuning protein structure and function (Müller et al., 2013). However, this experiment was carried out on protein enzymes, and it should also be demonstrated for ribozymes with amino acid coenzymes. Doig also discusses why some α-amino acids are not employed (Doig, 2017): for example they have too high energetic cost, or would not offer much novelty above the amino acids already present.

Actually, the main problem with code enlargement scenarios is the unknown period in which proteins had evolved. They might have a significant impact on selective forces. If some late amino acids were added after translation, then folding requirement could have been important and played a role in selection of some amino acids over others. On the other hand, the first few amino acids were not selected based on their future role in folded peptides. As there is no consensus on the order of amino acid assignment to the code, this requirement can be crucial. For example, if we accept that glycine, due to its availability on primordial Earth, was among the first amino acids to be incorporated into the genetic code, then its ability to freely rotate is an exaption (*sensu* (Gould and Vrba, 1982)). But if amino acids were selected in a metabolically rich RNA world based on their contribution to metabolism (for example, catalysis as proposed in (Kun et al., 2007)), then glycine had to be included later because of its flexibility. We do not yet know the order in which amino acids entered the code.

The error minimization scenario

Mutations can change the nucleotide sequence of genes. Due to the structure of the genetic code some of these changes, especially those affecting the 3rd position do not change the amino acid sequence of the coded polypeptides (same-sense mutations). Even miss-sense mutations do not by necessity change the protein as amino acids with similar physicochemical characteristic can substitute each other to some degree. Therefore there was an idea that the structure of the genetic code is such that point mutations minimize the physicochemical change of the coded amino acid.

The error minimization theory gained quite some credit when it was shown that the standard genetic code is quite robust against mutations (Ardell, 1998; Freeland et al., 2003; Gilis et al., 2001; Haig and Hurst, 1991; Kumar and Saini, 2016; Novozhilov and Koonin, 2009). But in none of these investigations did the standard genetic code emerge as the most resistant to mutations, and better codes can be designed from various points of views (Kuruoglu and Arndt, 2017). One can argue that any further optimization of the genetic code would have negligible benefit, and would have – as stated by Crick – a high cost. Was there a period for which the genetic code underwent optimization?

Let us assume for a moment that the code was extensively optimized by evolution by swapping codon assignment to arrive at a more optimized code. This mechanism is at the heart of the error minimalization scenario. In this case the only discernible pattern of the

standard genetic code would be error minimization. All other patterns that are suggestive of its origin or its extension would have been erased by the reassignment of the codon – amino acids pairs. But these patterns are there (Knight et al., 1999; Kun et al., 2007; Taylor and Coates, 1989), for example biosynthetically close amino acids have similar codons (see the coevolution theory above). Based on this Di Giulio also argues against the physicochemical theories (Di Giulio, 2017).

Modelling the evolution of the code as opposed to just analysing its current state also casts doubt to the physicochemical theory. Sengupta and co-workers conclude that (Bandhu et al., 2013): "the code evolution trajectory is possibly affected by extraneous factors and cannot be explained solely by natural selection between competing codes distinguished by differences in the level of physicochemical optimization." Furthermore, it was shown that the built up of the genetic code by the coevolution theory or the 2-1-3 model results is error minimization comparable and sometimes even better than in standard genetic code (Massey, 2008, 2016). Thus the error minimization feature of the genetic code could be a by-product of its evolution based on other mechanisms.

We do not want to belittle the error minimization capacity of the genetic code. It is important. But it might not have been played out by extensive swapping of the assignment during the evolution of the code. The error minimalization capacity selected for, but emerged as a consequence of how the genetic code evolved.

The frozen accident

Francis Crick's frozen accident theory often cited as one advocating that the assignment of codons and amino acids are random. It is clearly not so. The block structure of genetic code is a nonrandom pattern. We think the emphasis is on the frozen part, and not so much on the accident. To cite a sentence about the genetic code from the seminal paper: "at present time any change would be lethal, or at least very strongly selected against" (Crick, 1968). Once translation evolved and there were many polypeptides, the genetic code could not change much afterwards. Since the last universal common ancestor, which could have lived some 3+ billions of years ago, the genetic code changed very little (Keeling, 2016; Knight et al., 2001). There are some reassignments of codons, and there are two amino acids (selenocysteine and pyrrolysine) which began to enter the code. And that is it. The genetic code is indeed frozen, albeit not completely.

Chance (cf. accident) probably played some or even considerable role in the evolution of the genetic code (Koonin, 2017). There is chance element is the choice of the particular 20 amino acids employed. For example, „the selection of isoleucine over alloisoleucine seems to be chance" (Doig, 2017).

If the genetic code was frozen when translated polypeptides become widespread, then we can infer the complexity of the genetic code prior to this point. Methionine and tryptophan, each having only a single codon, could have been the last amino acids to enter the standard genetic code. They might have entered the genetic code after it has mostly frozen. Similarly, selenocysteine and pyrrolysine are entering the genetic code of some organisms. Thus the rest of the 20 amino acids, 18 of them (methionine and tryptophan could have entered afterward),

could have already been established in the genetic code before most of the ribozymes were replaced by protein enzymes.

There is another theory which corroborates the aforementioned idea: the genetic code could have been frozen because there are only a finite number of discrimination positions on the tRNAs (Ribas de Pouplana et al., 2017; Saint-Léger et al., 2016). Here we need to mention, that sometimes identity element can be the absence of certain nucleotides at certain places (negative identity determinants), and not the presence of it as demonstrated for the class I and class II tRNA identity elements (Jakó et al., 2007). Discriminator / identity elements are important as the genetic code is "known" by the aminoacyl tRNA synthetases, the enzymes that attach the amino acid to the cognate tRNA. These enzymes do not always discriminate based on the anticodon, thus the assignment of amino acid to the tRNA, and the attachment of the tRNA anticodon loop and the mRNA could be independent. This leads to the idea of a "second genetic code" (de Duve, 1988).

The frozen accident theory is the only one that says anything about why there are 20 amino acids: because the code froze before the inclusion of more. On the other hand, the genetic code could become frozen because the 20 (or 18 if Met and Trp entered later) amino acid coded already covers the physic-chemical range for charge, size and hydrophobicity of α-amino acids of prebiotic importance and ones that can be reached through our current metabolism (i.e. intermediates included) (Philip and Freeland, 2011). Furthermore, from the folding point of view the set of 20 amino acids employed is quite good (Doig, 2017).

The impasse and the challenge

Papers about the origin of the genetic code list the same main ideas (see above), many of which were established several decades ago. Scientists are still debating which of them or the combination of which of them is true.

The most well studied scenario is the error minimization scenario. Over the years it was thus shown that this scenario is not important for the evolution of the genetic code. One hypothesis is out. Nowadays theorists try to combine the hypotheses into a coherent adaptation story (like we tried in this paper). This is good, as integration of knowledge is one way to further the scientific enterprise. For example, the stereochemical, the coding coenzyme handle and the coevolution hypotheses combine well together (Szathmáry and Zintzaras, 1992; Taylor and Coates, 1989). But there is a disagreement on the order of amino acids to enter the code. The coevolution theory starts with simple amino acids in the code, but the stereochemical hypothesis, which should be the basis of the first assignment to codons, is mostly demonstrated for supposedly late amino acids. The lack of data for Asp is especially troubling. The coevolution hypothesis assumes that the same pathways produced the amino acids as in contemporary bacteria. As there are alternative pathways for some amino acids, there could have been alternative pathways in the RNA world. Furthermore, we need to prove that amino acids can be produced by ribozymes, i.e. all steps in the pathway can be catalysed by RNA enzymes or RNA enzymes employing amino acid cofactors that are already producible or present.

To be honest, we are stuck. The existing theories each capture some aspect of the origin of the genetic code, but they still contain a lot of assumptions that could be cleared by

experiments. New theories are either slight variation of old theories (science mostly advance incrementally) or theories that make little impact on the literature (which is unfortunate, there could be a lot of good idea out there to be discovered). So there is an impasse. Most papers on the origin of the genetic code are reviews (like this) and not original research. It seems that the field has been stalled. We strongly urge empiricist to conduct the experiments we proposed here in order to overcome this impasse and go on with the challenging task of solving the "notoriously difficult problem" of the origin of the genetic code and translation.

Acknowledgement

We are grateful for Eörs Szathmáry for his comment on earlier draft of the manuscript. Financial support has been provided by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013)/ERC grant agreement no [294332]; the Hungarian National Research, Development and Innovation Office under the grant agreement no NKFI K 119347; and by GINOP 2.3.2-15-2016-00057. This work was carried out as part of EU COST action CM1304 "Emergence and Evolution of Complex Chemical Systems".

References

Adamala, K., Szostak, J.W., 2013. Competition between model protocells driven by an encapsulated catalyst. Nature Chemistry 5, 495–501.

Agresti, J.J., Kelly, B.T., Jaschke, A., Griffiths, A.D., 2005. Selection of ribozymes that catalyse multiple turnover Diels-Alder cycloadditions by using *in vitro* compartmentalization. PNAS 102, 16170–16175.

Ardell, D.H., 1998. On error minimization in a sequential origin of the standard genetic code. J. Mol. Evol. 47, 1–13.

Bada, J.L., 2013. New insights into prebiotic chemistry from Stanley Miller's spark discharge experiments. Chem. Soc. Rev. 42, 2186–2196.

Bandhu, A.V., Aggarwal, N., Sengupta, S., 2013. Revisiting the physico-chemical hypothesis of code origin: An analysis based on code-sequence coevolution in a finite population. Origins Life Evol. Biosphere 43, 465–489.

Barbieri, M., 2015. Evolution of the genetic code: The ribosome-oriented model. Biological Theory 10, 301–310.

Bartlett, G.J., Porter, C.T., Borkakoti, N., Thornton, J.M., 2002. Analysis of catalytic residues in enzyme active sites. J. Mol. Biol. 324, 105–121.

Berman, H., Henrick, K., Nakamura, H., 2003. Announcing the worldwide Protein Data Bank. Nat. Struct. Mol. Biol. 10, 980.

Bernhardt, H.S., Patrick, W.M., 2014. Genetic code evolution started with the incorporation of glycine, followed by other small hydrophilic amino acids. J. Mol. Evol. 78, 307-309.

Bernhardt, H.S., Tate, W.P., 2008. Evidence from glycine transfer RNA of a frozen accident at the dawn of the genetic code. Biology Direct 3, 53.

Brooks, D.J., Fresco, J.R., Lesk, A.M., Singh, M., 2002. Evolution of amino acid frequencies in proteins over deep time: Inferred order of introduction of amino acids into the genetic code. Mol. Biol. Evol. 19, 1645–1655.

Connell, G.J., Illangesekare, M., Yarus, M., 1993. Three small ribooligonucleotides with specific arginine sites. Biochemistry 32, 5497–5502.

Crick, F.H.C., 1968. The origin of the genetic code. J. Mol. Biol. 38, 367–379.

Crick, F.H.C., Brenner, S.E., Klug, A., Pieczenik, G., 1976. A speculation on the origin of protein synthesis. Origin of Life 7, 389–397.

de Duve, C., 1988. The second genetic code. Nature 333, 117–118.

de Vladar, H., 2012. Amino acid fermentation at the origin of the genetic code. Biology Direct 7, 6.

Di Giulio, M., 2008. An extension of the coevolution theory of the origin of the genetic code. Biology Direct 3, 37.

Di Giulio, M., 2017. Some pungent arguments against the physico-chemical theories of the origin of the genetic code and corroborating the coevolution theory. J. Theor. Biol. 414, 1–4.

Doig, A.J., 2017. Frozen, but no accident – why the 20 standard amino acids were selected. The FEBS Journal 284, 1296–1305.

Eigen, M., Winkler-Oswatitsch, R., 1981. Transfer-RNA, an early gene? Naturwissenschaften 68, 282–292.

Freeland, S.J., Wu, T., Keulmann, N., 2003. The case for an error minimizing standard genetic code. Origins Life Evol. Biosphere 33, 457–477.

Furnham, N., Holliday, G.L., de Beer, T.A.P., Jacobsen, J.O.B., Pearson, W.R., Thornton, J.M., 2014. The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. Nucleic Acids Res. 42, D485–D489.

Gil, R., Silva, F.J., Peretó, J., Moya, A., 2004. Determination of the core of a minimal bacterial gene set. Microbiol. Mol. Biol. Rev. 68, 518–537.

Gilbert, W., 1986. Origin of life: the RNA world. Nature 319, 618.

Gilis, D., Massar, S., Cerf, N.J., Rooman, M., 2001. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. Genome Biol. 2, research0049.0041–research0049.0012.

Gorlero, M., Wieczorek, R., Adamala, K., Giorgi, A., Schininà, M.E., Stano, P., Luisi, P.L., 2009. Ser-His catalyses the formation of peptides and PNAs. FEBS Lett. 583, 153–156.

Gould, S.J., Vrba, E.S., 1982. Exaptation — a missing term in the science of form. Paleobiology 8, 4–15.

Griffiths, A.D., Tawfik, D.S., 2006. Miniaturising the laboratory in emulsion droplets. Trends Biotechnol. 24, 395–402.

Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., Altman, S., 1983. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell 35, 849–857.

Gulik , P., 2015. On the origin of sequence. Life 5, 1629.

Haig, D., Hurst, L.D., 1991. A quantitative measure of error minimization in the genetic code. J. Mol. Evol. 33, 412–417.

Higgs, P., 2009. A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. Biology Direct 4, 16.

Higgs, P.G., Pudritz, R.E., 2009. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code Astrobiology 9, 483–490.

Ikehara, K., 2016. Evolutionary steps in the emergence of life deduced from the bottom-up approach and GADV hypothesis (top-down approach). Life 6, 6.

Ikehara, K., Omori, Y., Arai, R., Hirose, A., 2002. A novel theory on the origin of the genetic code: A GNC-SNS hypothesis. J. Mol. Evol. 54, 530–538.

Illangasekare, M., Yarus, M., 2002. Phenylalanine-binding RNAs and genetic code evolution. J. Mol. Evol. 54, 298–311.

Jadhav, V.R., Yarus, M., 2002. Coenzymes as coribozymes. Biochimie 84, 877–888.

Jakó, É., Ittzés, P., Szenes, Á., Kun, Á., Szathmáry, E., Pál, G., 2007. *In silico* detection of tRNA sequence features characteristic to aminoacyl-tRNA synthetase class membership. Nucleic Acids Res. 35, 5593–5609.

Janas, T., Widmann, J.J., Knight, R., Yarus, M., 2010. Simple, recurring RNA binding sites for L-arginine. RNA 16, 805–816.

Jauker, M., Griesser, H., Richert, C., 2015. Spontaneous formation of RNA strands, peptidyl RNA, and cofactors. Angew. Chem. Int. Ed. 54, 14564–14569.

Johnston, W.K., Unrau, P.J., Lawrence, M.S., Glasen, M.E., Bartel, D.P., 2001. RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. Science 292, 1319–1325.

Keeling, Patrick J., 2016. Genomics: Evolution of the genetic code. Curr. Biol. 26, R851–R853.

Knight, R.D., Freeland, S.J., Landweber, L.F., 1999. Selection, history and chemistry: the three faces of the genetic code. Trends Biochem. Sci 24, 241–247.

Knight, R.D., Freeland, S.J., Landweber, L.F., 2001. Rewiring the keyboard: evolvability of the genetic code. Nat. Rev. Genet. 2, 49-58.

Koonin, E., 2017. Frozen accident pushing 50: Stereochemistry, expansion, and chance in the evolution of the genetic code. Life 7, 22.

Kruger, K., Grabowski, P., Zaug, A.J., Sands, J., Gottschling, D.E., Cech, T.R., 1982. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. Cell 31, 147–157.

Krupkin, M., Matzov, D., Tang, H., Metz, M., Kalaora, R., Belousoff, M.J., Zimmerman, E., Bashan, A., Yonath, A., 2011. A vestige of a prebiotic bonding machine is functioning within the contemporary ribosome. Philos. Trans. R. Soc. Lond., Ser. B: Biol. Sci. 366, 2972–2978.

Kumar, B., Saini, S., 2016. Analysis of the optimality of the standard genetic code. Mol. BioSys. 12, 2642–2651.

Kun, Á., Pongor, S., Jordán, F., Szathmáry, E., 2007. Catalytic propensity of amino acids and the origins of the genetic code and proteins, In: Barbieri, M. (Ed.), The Codes of Life: The Rules of Macroevolution. Springer, pp. 39–58.

Kun, Á., Szilágyi, A., Könnyű, B., Boza, G., Zachár, I., Szathmáry, E., 2015. The dynamics of the RNA world: Insights and challenges. Ann. N.Y. Acad. Sci. 1341, 75–95.

Kuruoglu, E.E., Arndt, P.F., 2017. The information capacity of the genetic code: Is the natural code optimal? J. Theor. Biol. 419, 227–237.

Lazcano, A., 2010. Historical development of origins research. Cold Spring Harb. Perspect. Biol. 2, a002089.

Legiewicz, M., Yarus, M., 2005. A more complex isoleucine aptamer with a cognate triplet. J. Biol. Chem. 280, 19815–19822.

Longo, L.M., Blaber, M., 2012. Protein design at the interface of the pre-biotic and biotic worlds. Arch. Biochem. Biophys. 526, 16–21.

Lozupone, C., Changayil, S., Majerfeld, I., Yarus, M., 2003. Selection of the simplest RNA that binds isoleucine. RNA 9, 1315–1322.

Majerfeld, I., Puthenvedu, D., Yarus, M., 2005. RNA affinity for molecular L-histidine; genetic code origins. J. Mol. Evol. 61, 226–235.

Majerfeld, I., Yarus, M., 1998. Isoleucine:RNA sites with associated coding sequences. RNA 4, 471–478.

Majerfeld, I., Yarus, M., 2005. A diminutive and specific RNA binding site for L-tryptophan. Nucleic Acids Res. 33, 5482–5493.

Mannironi, C., Scerch, C., Fruscoloni, P., Tocchini-Valentini, G.P., 2000. Molecular recognition of amino acids by RNA aptamers: The evolution into an L-tyrosine binder of a dopamine-binding RNA motif. RNA 6, 520–527.

Massey, S.E., 2006. A sequential "2-1-3" model of genetic code evolution that explains codon constraints. J. Mol. Evol. 62, 809–810.

Massey, S.E., 2008. A neutral origin for error minimization in the genetic code. J. Mol. Evol. 67, 510.

Massey, S.E., 2016. The neutral emergence of error minimized genetic codes superior to the standard genetic code. J. Theor. Biol. 408, 237–242.

Matsumura, S., Kun, Á., Ryckelynck, M., Coldren, F., Szilágyi, A., Jossinet, F., Rick, C., Nghe, P., Szathmáry, E., Griffiths, A.D., 2016. Transient compartmentalization of RNA replicators prevents extinction due to parasites. Science 354, 1293–1296.

Miller, O.J., Bernath, K., Agresti, J.J., Amitai, G., Kelly, B.T., Mastrobattista, E., Taly, V., Magdassi, S., Tawfik, D.S., Griffiths, A.D., 2006. Directed evolution by in vitro compartmentalization. Nat. Methods 3, 561–570.

Morris, C.E., 2002. How did cells get their size? The Anatomical Record 268, 239–251.

Müller, M.M., Allison, J.R., Hongdilokkul, N., Gaillon, L., Kast, P., van Gunsteren, W.F., Marlière, P., Hilvert, D., 2013. Directed evolution of a model primordial enzyme provides insights into the development of the genetic code. PLoS Genet. 9, e1003187.

Noller, H.F., 2004. The driving force for molecular evolution of translation. RNA 10, 1833–1837.

Novozhilov, A.S., Koonin, E.V., 2009. Exceptional error minimization in putative primordial genetic codes. Biology Direct 4, 44.

Philip, G.K., Freeland, S.J., 2011. Did evolution select a nonrandom "alphabet" of amino acids? Astrobiology 11, 235–240.

Porter, C.T., Bartlett, G.J., Thornton, J.M., 2004. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res. 32, D129–D133.

Reeve, H.K., Sherman, P.W., 1993. Adaptation and the goals of evolutionary research. Q. Rev. Biol. 68, 1–32.

Ribas de Pouplana, L., Torres, A., Rafels-Ybern, À., 2017. What froze the genetic code? Life 7, 14.

Root-Bernstein, R., 2010. Experimental test of L- and D-amino acid binding to L- and D-codons suggests that homochirality and codon directionality emerged with the genetic code. Symmetry 2, 1180.

Roth, A., Breaker, R.R., 1998. An amino acid as a cofactor for a catalytic polynucleotid. PNAS 95, 6027–6031.

Saint-Léger, A., Bello, C., Dans, P.D., Torres, A.G., Novoa, E.M., Camacho, N., Orozco, M., Kondrashov, F.A., Ribas de Pouplana, L., 2016. Saturation of recognition elements blocks evolution of new tRNA identities. Science Advances 2, e1501860.

Saran, D., Frank, J., Burke, D.H., 2003. The tyranny of adenosine recognition among RNA aptamers to coenzyme A. BMC Evol. Biol. 3, 26.

Shimizu, M., 1995. Specific aminoacylation of C4N hairpin RNAs with the cognate aminoacyl-adenylates in the presence of a dipeptide: Origin of the genetic code. The Journal of Biochemistry 117, 23–26.

Shimizu, M., 1996. Detection of the peptidyltransferase activity of a dipeptide, alanylhistidine, in the absence of ribosomes. The Journal of Biochemistry 119, 832–834.

Sober, E., 1984. The Nature of Selection: Evolutionary Theory in Philosophical Focus. Bradford/MIT Press.

Szathmáry, E., 1990. Useful coding before translation: the coding coenzymes handle hypothesis for the origin of the genetic code., In: Lukács, B., Bérczi, S., Molnár, I., Paál, G. (Eds.), Evolution: from Cosmogenesis to Biogenesis. KFKI-1990-50/C, Budapest, pp. 77–83.

Szathmáry, E., 1993. Coding coenzyme handles: a hypothesis for the origin of the genetic code. PNAS 90, 9916–9920.

Szathmáry, E., 1996. Coding coenzyme handles and the origin of the genetic code., In: Müller, A., Dress, A., Vögtle, F. (Eds.), From Simplicity to Complexity in Chemistry -- and Beyond. Part I. Vieweg, Braunschweig, pp. 33–41.

Szathmáry, E., 1999. The origin of the genetic code – amino acids as cofactors in an RNA world. Trends Genet. 15, 223–229.

Szathmáry, E., Zintzaras, E., 1992. A statistical test of hypotheses on the organization and origin of the genetic code. J. Mol. Evol. 35, 185–189.

Tamura, K., 2015. Beyond the frozen accident: Glycine assignment in the genetic code. J. Mol. Evol. 81, 69-71.

Taylor, F.J.R., Coates, D., 1989. The code within the codons. BioSyst. 22, 177–187.

Trifonov, E.N., 2000. Consensus temporal order of amino acids and evolution of the triplet code. Gene 261, 139–151.

Trifonov, E.N., 2004. The triplet code from first principles. J. Biomol. Struct. Dyn. 22, 1–11.

Tsukiji, S., Pattnaik, S.B., Suga, H., 2003. An alcohol dehydrogenase ribozyme. Nat. Struct. Mol. Biol. 10, 713–717.

Turk-MacLeod, R.M., Puthenvedu, D., Majerfeld, I., Yarus, M., 2012. The plausibility of RNA-templated peptides: Simultaneous RNA affinity for adjacent peptide side chains. J. Mol. Evol. 74, 217–225.

Turk, R.M., Chumachenko, N.V., Yarus, M., 2010. Multiple translational products from a five-nucleotide ribozyme. PNAS 107, 4585–4589.

Weber, A.L., Pizzarello, S., 2006. The peptide-catalyzed stereospecific synthesis of tetroses: A possible model for prebiotic molecular evolution. Proceedings of the National Academy of Sciences 103, 12713–12717.

West-Eberhard, M.J., 1992. Adaptation: Curreent usages, In: Keller, E.F., Lloyd, E.A. (Eds.), Keywords in Evolutionary Biology. Harvard University Press, Cambridge, pp. 13–18.

White, H.B., 1976. Coenzymes as fossils of an earlier metabolic state. J. Mol. Evol. 7, 101–104.

Widman, J., di Giulio, M., Yarus, M., Knight, R., 2005. tRNA creation by hairpin duplication. J. Mol. Evol. 61, 524–530.

Woese, C.R., Dugre, D.H., Dugre, S.A., Kondo, M., Saxinger, W.C., 1966a. On the fundamental nature and evolution of the genetic code. Cold Spring Harbor Symp. Quant. Biol. 31, 723–736.

Woese, C.R., Dugre, D.H., Saxinger, W.C., Dugre, S.A., 1966b. The molecular basis for the genetic code. PNAS 55, 966–974.

Wong, J.-F., 2007. Question 6: Coevolution theory of the genetic code: A proven theory. Origins Life Evol. Biosphere 37, 403–408.

Wong, J., Ng, S.-K., Mat, W.-K., Hu, T., Xue, H., 2016. Coevolution theory of the genetic code at age forty: Pathway to translation and synthetic life. Life 6, 12.

Wong, J.T.-F., 2005. Coevolution theory of the genetic code at age thirty. Bioessays 27, 416–425.

Wong, J.T.F., 1975. A co-evolution theory of the genetic code. PNAS 72, 1909–1912.

Yao, C., Moretti, J.E., Struss, P.E., Spall, J.A., Müller, U.F., 2011. Arginine cofactors on the polymerase ribozyme. PLoS ONE 6, e25030.

Yarus, M., 1988. A specific amino acid binding site composed of RNA. Science 240, 1751–1758.

Yarus, M., 2017. The genetic code and RNA-amino acid affinities. Life 7, 13.

Yarus, M., Caporaso, J.G., Knight, R., 2005. Origins of the genetic code: the escaped triplet theory. Annu. Rev. Biochem 74, 179–198.

Yarus, M., Widmann, J.J., Knight, R., 2009. RNA-amino acid binding: a stereochemical era for the genetic code. J. Mol. Evol. 69, 406–429.

Zachar, I., Szathmáry, E., 2010. A New Replicator: a theoretical framework for analysing replication. BMC Biol. 8, 21.