Reciprocal Nucleopeptides as the Ancestral Darwinian Self-Replicator

Eleanor F. Banwell,[†,1] Bernard M.A.G. Piette,[†,2] Anne Taormina,[2] and Jonathan G. Heddle*[,1,3]

[1]Heddle Initiative Research Unit, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

[2]Department for Mathematical Sciences, Durham University, Durham, United Kingdom

[3]Bionanoscience and Biochemistry Laboratory, Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: jonathan.heddle@uj.edu.pl.

Associate editor: Nicole Perna

Abstract

Even the simplest organisms are too complex to have spontaneously arisen fully formed, yet precursors to first life must have emerged ab initio from their environment. A watershed event was the appearance of the first entity capable of evolution: the Initial Darwinian Ancestor. Here, we suggest that nucleopeptide reciprocal replicators could have carried out this important role and contend that this is the simplest way to explain extant replication systems in a mathematically consistent way. We propose short nucleic acid templates on which amino-acylated adapters assembled. Spatial localization drives peptide ligation from activated precursors to generate phosphodiester-bond-catalytic peptides. Comprising autocatalytic protein and nucleic acid sequences, this dynamical system links and unifies several previous hypotheses and provides a plausible model for the emergence of DNA and the operational code.

Key words: **Initial Darwinian Ancestor, abiogenesis, RNA world, protein world, nucleopeptide replicator, reciprocal replicator, polymerase, ribosome, evolution, early earth, hypercycle.**

Article

Introduction

In contrast to our good understanding of more recent evolution, we still lack a coherent and robust theory that adequately explains the initial appearance of life on Earth (abiogenesis). In order to be complete, an abiogenic theory must describe a path from simple molecules to the Last Universal Common Ancestor (LUCA), requiring only a gradual increase in complexity.

The watershed event in abiogenesis was the emergence of the Initial Darwinian Ancestor (IDA): the first self-replicator (ignoring dead ends) and ancestral to all life on Earth (Yarus 2011). Following the insights of von Neumann, who proposed the kinematic model of self-replication (Kemeny 1955), necessary features of such a replicator are: Storage of the information for how to build a replicator; a processor to interpret information and select parts; an instance of the replicator.

In order to be viable, any proposal for the IDA's structure must fit with spontaneous emergence from prebiotic geochemistry and principles of self-replication. Currently, the most dominant abiogenesis theory is the "RNA world," which posits that the IDA was a self-replicating ribozyme, that is, an RNA-dependent RNA polymerase (Cech 2012). Although popular, this theory has problems (Kurland 2010). For example, while it is plausible that molecules with the necessary replication characteristics can exist, length requirements seem to make their spontaneous emergence from the primordial milieu unlikely, nor does the RNA world explain the appearance of the operational code (Noller 2012; Robertson and Joyce 2012). Furthermore, it invokes three exchanges of function between RNA and other molecules to explain the coupling of polynucleotide and protein biosynthesis, namely transfer of information storage capability to DNA and polymerase activity to protein as well as gain of peptide synthesis ability. This presents a situation in which no extant molecule continues in the role it initially held. Others have posited peptide and nucleopeptide worlds as solutions.

The peptide world theory proposes a spontaneously occurring self-replicating peptide with RNA synthesis, DNA and the operational code appearing later, and possible self-replicating mechanisms of peptides have been explored (Fox and Harada 1958; Lee et al. 1996). Nucleopeptide theories require that the replicator consist of both peptides and nucleic acids and may involve their covalent linkage or (as in our proposal) noncovalent conjugation. Covalently linked nucleopeptides include nucleobase-containing peptides such as PNA which has been mooted as a possible precursor to the RNA world (Miller 1997) and possible RNA-interacting nucleo-∈-peptides have been synthesized (Roviello et al. 2009; Nelson et al. 2000). Both the peptide world and nucleopeptide theories consist of single molecular classes and therefore suffer the same exchange of function problems as the RNA-world theory. To the best of our knowledge, no single theory has emerged that parsimoniously answers the biggest questions.

Here, we build on several foregoing concepts to propose an alternative theory based around a nucleopeptide reciprocal replicator that uses its polynucleotide and peptide

components according to their strengths, thus avoiding the need to explain later exchange of function and coupling. We advocate a view of the IDA resulting from a biochemical system which we describe as a dynamical system, that is, a system of equations describing the changes that occur over time in the self-replicator presented here, and we demonstrate that such an entity is both mathematically consistent and complies with all the logical requirements for life. While necessarily wide in view we hope that this work will provide a useful framework for further investigation of this fundamental question.

Model and Results

Solving the Chicken and Egg Problem

Given that any IDA must have been able to replicate in order to evolve, extant cellular replication machinery is an obvious source of clues to its identity. Common ancestry means that features shared by all life were part of LUCA. By examining the common replication components present in LUCA, and then extrapolating further back to their simplest form, it is possible to reach a pre-LUCA, irreducibly complex, core replicator (fig. 1).

We see that in all cells, the required functions of a replicator are not carried out by a single molecule or even a single class of molecules, rather they are performed variously by nucleic acids (DNA, RNA) and proteins. When viewed by molecular class, the replicator has two components and is reciprocal in nature: polynucleotides rely on proteins for their polymerization and vice versa. The question of which arose first is a chicken and egg conundrum that has dogged the field since the replication mechanisms were first elucidated (Giri and Jain 2012). In this work, we suggest that, consistent with common ancestry and in contrast with the RNA world theory, the earliest replicator was a two—rather than a one—component system, composed of peptide and nucleic acids.

Assumptions of the Model

We postulate that, in a nucleopeptide reciprocal replicator, the use of each component according to its strengths could deliver a viable IDA more compatible with evolution to LUCA replication machinery. Although seemingly more complex than an individual replicating molecule, the resulting unified abiogenesis theory answers many hard questions and is ultimately more parsimonious. The model does not consider in detail the chemistry of how the building blocks that constitute the IDA (short peptides and nucleic acids) came about as these details are covered in the cited literature (see for example, Saladino et al. 2012; Patel et al. 2015; Da Silva et al. 2015; Leman et al. 2004; Liu et al. 2014; Martin et al. 2008). Rather, we concentrate on the important question of the mathematical validity of the IDA in terms of its ability to sustainably self-replicate, without which it would not be a valid system. In constructing our model, we make the following assumptions:

(i) The existence of random sequences of short strands of mixed nucleic acids (XNA) likely consisting of ribonucleotides, deoxyribonucleotides and possibly other building blocks able to polymerize with nucleotide chains, as well as the existence of random amino acids and short peptides produced abiotically.

For this first assumption we have supposed a pool of interacting amino acids, nucleotides and related small molecules as well as a supply of metal ions, other inorganic catalysts and energy. The precise understanding of the "metabolic" reactions in which these precursor building blocks were formed is in itself an extremely important question but is not considered here as a number of potential early earth conditions and reaction pathways resulting in these outcomes have already been proposed, including the formamide reaction (Saladino et al. 2012) and cyanosulfidic chemistries (Patel et al. 2015). Recent experimental models of alkaline hydrothermal vents have even succeeded in producing various organic molecules including ribose and deoxyribose (Herschy et al. 2014). Pools

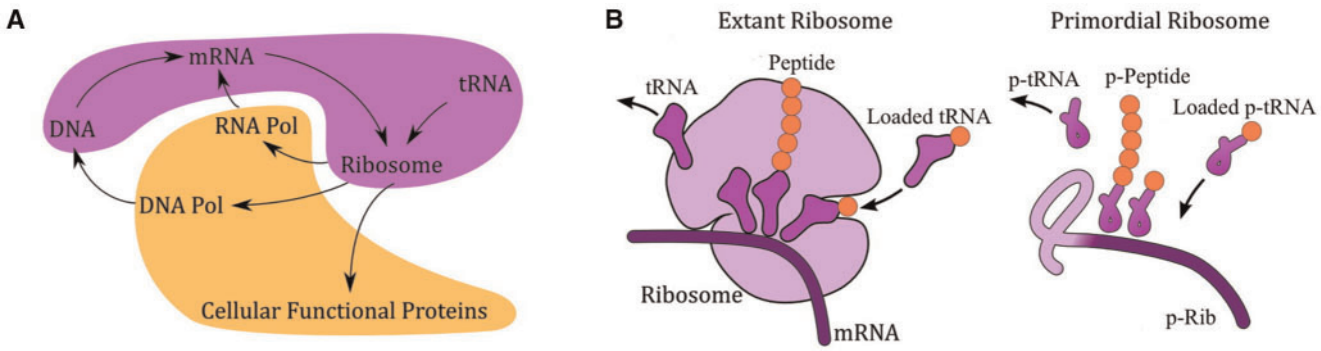

FIG. 1. Replication schemes. (*a*) This simplified cellular replication schematic is common to all life today and likely reflects the ancestral form present in LUCA. Shading by molecule type (purple for nucleic acid and orange for protein), reveals a reciprocal nucleopeptide replicator. Although the ribosome is a large nucleoprotein complex, the catalytic centre has been shown to be a ribozyme (Moore and Steitz 2003) and so it is shaded purple in this scheme. (*b*) Comparison of the method of action of the extant ribosome with the proposed primordial analogue (components are shaded like for like). Today, tRNA molecules (mid purple) loaded with amino acids (orange) bind the mRNA (dark purple) in the ribosome (light purple), which co-ordinates and catalyses the peptidyl-transferase reaction. Although the present day modus operandi is regulated via far more complex interactions than the primordial version, the two schemes are fundamentally similar. Mixed nucleic acid structures, one performing a dual function as primordial mRNA and primordial ribosome (p-Rib) and a second functioning as a primordial tRNA (p-tRNA), provide a system wherein the former structure templates amino acid-loaded molecules of the latter.

of pure molecules are unlikely; instead, mixtures would likely have comprised standard and nonstandard amino acids as well as XNAs with mixed backbone architectures, being, in their simplest forms, mixtures of deoxy- and ribonucleotides (Trevino et al. 2011; Pinheiro et al. 2012) with other building blocks being possible. For simplicity we sometimes refer to XNAs as "polynucleotides." Such conditions would be conducive to the occasional spontaneous covalent attachment of nucleotides to each other to form longer polymer chains (Da Silva et al. 2015).

(ii) The existence of abiotically aminoacylated short XNA strands (primordial tRNAs (p-tRNAs))

The second assumption is potentially troubling as amino acid activation is slow and thermodynamically unfavorable. However, amino acylation has been investigated in some detail and has been shown to be possible abiotically including, in some cases, the abiotic production of activated amino acids (Illangasekare et al. 1995; Leman et al. 2004; Giel-Pietraszuk and Barciszewski 2006; Lehmann et al. 2007; Turk et al. 2010; Liu et al. 2014). A pool of activated amino acids allows us to presume a fast rate of charging of p-tRNAs meaning that we can assume that the rate of charged p-tRNA formation is proportional to the concentration of free amino acids. Taken together these data suggest that multiple small amino-acylated tRNA-like primordial XNAs could have arisen. Though likely being XNA in nature, we refer to them as p-tRNA, reflecting their function. A similar nomenclature applies to p-Rib and p-mRNA.

(iii) Conditions that allow a codon/anticodon interaction between two or more charged p-tRNA for sufficient time and appropriate geometry to allow peptide bond formation, that is, the functionality of a primordial ribosome (p-Rib)

Our proposed p-Rib is an extreme simplification of the functionality of both the present day ribosome and mRNA (fig. 1). Initially, the p-Rib need only have been a (close to) linear assembly template for the p-tRNAs to facilitate the peptidyl transferase reaction through an increase in local concentration. This mechanism is simple enough to emerge spontaneously and matches exactly the fundamental action of the extant ribosome (fig. 2). The idea that a p-Rib may have an internal template rather than separate mRNA molecules and that an RNA strand could act as a way to bring charged tRNAs together has previously been suggested (Schimmel and Henderson 1994; Wolf and Koonin 2007; Morgens 2013) and is known as an "entropy trap" (Sievers et al. 2004; Ruiz-Mirazo et al. 2014). The concept has been demonstrated to be experimentally viable (Tamura and Schimmel 2003) although in the latter case it is the primordial ribosomal rRNA strand itself that provides one of the two reacting amino acids.

A functional operational system requires preferential charging of particular p-tRNAs to specific amino acids. Although there is evidence for such relationships in the stereochemical theory (Woese 1965; Yarus et al. 2009), so far unequivocal proof has been elusive (Yarus et al. 2005; Koonin and Novozhilov 2009). However, there is sufficient evidence to suggest at least a separation along grounds of hydrophobicity and charge using just a two-base codon (Knight and Landweber 2000; Biro et al. 2003; Rodin et al. 2011). Furthermore only a reduced set of amino acids (Angyan et al. 2014)—possibly as few as four (Ikehara 2002)—need to have been provided in this way. The "statistical protein" hypothesis proposes that such a weak separation may have been sufficient to produce populations of active peptides (Ikehara 2005; Vetsigian et al. 2006). Such "primordial polymerases" (p-Pol) need only have been small (see below) and spontaneous emergence of a template coding loosely for such a sequence seems plausible. The failure rate of such syntheses would be high but a p-Rib using the outlined primordial operational code to produce statistical p-Pol peptides could have been accurate enough to ensure its own survival.

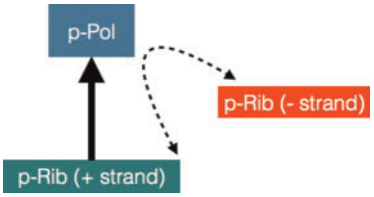

Fig. 2. Models of primitive polymerization reactions. An XNA strand can function like a primordial ribosome (p-Rib) whereby one strand (+ strand) can template the production of a primordial polymerase (p-Pol) as indicated by the solid arrow. The action of this p-Pol is represented by the double-headed dotted arrow whereby it acts on the p-Rib (+ strand) to catalyze synthesis of the complementary sequence (− strand) and also on the − strand to produce more of the + strand.

(iv) The viability of a very short peptide sequence to function as an RNA-dependent RNA polymerase

Templated ligation is often proposed as a primordial self-replication mechanism, particularly for primitive replication of nucleic acid in RNA world type scenarios. However, these are associated with a number of problems as mentioned earlier. In addition, extant RNA/DNA synthesis proceeds via terminal elongation (Paul and Joyce 2004; Vidonne and Philp 2009). To be consistent with the mechanism present in LUCA and pre-LUCA, the p-Pol should, preferably, have used a similar process.

During templated ligation, a parent molecule binds and ligates short substrates that must then dissociate to allow further access, but the product has greater binding affinity than the substrates and dissociation is slow. This product inhibition results in parabolic growth and limits the usefulness of templated ligation for replication (Issac and Chmielewski 2002). Conversely, in 1D sliding (or more accurately jumping), the catalyst may dock anywhere along a linear substrate and then diffuse by "hops" randomly in either direction until it reaches the reaction site; a successful ligation reaction has little impact on binding affinity and leaves the catalyst proximal to the next site. For simplicity our model assumes a single binding event between p-Pol and p-Rib followed by multiple polymerization events. A p-Pol proceeding via 1D sliding could catalyze phosphodiester bond formation between nucleotides bound by Watson and Crick base-pairing to a complementary XNA strand. Because p-Pol activity would be independent of substrate length, a relatively small catalyst could have acted on XNAs of considerable size. From

inspection of present day polymerases such a peptide may have included sequences such as DxDGD and/or GDD known to be conserved in their active sites and consisting of the amino acids thought to be amongst the very earliest in life (Iyer et al. 2003; Koonin 1991).

In our simple system any such p-Pol must be very short to have any realistic chance of being produced by the primitive components described. We must therefore ask if there is evidence that small (e.g. <11 amino acid) peptides can have such a catalytic activity. Catalytic activity in general has been demonstrated for molecules as small as dipeptides (Kochavi et al. 1997). For polymerase activity in particular, it is known that randomly produced tripeptides can bind tightly and specifically to nucleotides (Schneider et al. 2000; McCleskey et al. 2003). We suggest that a small peptide could arise with the ability to bind divalent metal ions, p-Rib and incoming nucleotides. It is interesting to note that small peptides can assemble into large and complex structures (Bromley et al. 2008; Fletcher et al. 2013) with potentially sophisticated functionality: di- and tripeptides can self-assemble into larger nanotubes and intriguingly it has even been suggested that these structures could have acted as primitive RNA polymerases (Carny and Gazit 2005).

In summary, the essence of the model is that on geological timescales, short linear polynucleotides may have been sufficient to template similar base-pairing interactions to those seen in the modern ribosome with small amino-acylated adapters. Given that the majority of ribosome activity stems from accurate substrate positioning, such templating could be sufficient to catalyze peptide bond formation and to deliver phosphodiester-bond-catalytic peptides. As backbone ligation reactions are unrelated to polynucleotide sequence, these generated primordial enzymes could have acted on a large subset of the available nucleic acid substrates, in turn producing more polynucleotide templates and resulting in an autocatalytic system.

Mathematical Model

The IDA described above is attractive both for its simplicity and continuity with the existing mixed (protein/nucleic acid) replicator system in extant cells. However, the question remains as to whether such a system is mathematically consistent, could avoid collapse and instead become self-sustaining. The number of parameters and variables needed to analyze the system in its full complexity is such that one is led to consider simplified models which nevertheless capture essential features of interest. Here we consider a simple model of RNA–protein self-replication.

Constituents

The main constituents of the simplest model of XNA-protein self-replication considered here (see also figs. 1*b* and 2) are a pool of free nucleotides and amino acids, polypeptide chains—including a family of polymerases—and polynucleotide chains as well as p-tRNAs loaded with single amino acids.

We introduce some notations. Generically, we consider polymer chains Π made of *n types* of building blocks labeled $1, \ldots, n$. In our models, the polymer chains are polypeptides and polynucleotides, and the building blocks are amino acids and codons respectively. With a slight abuse of language, we call the number of constituents (building blocks) of a polymer chain its *length*. So hereafter, "lengths" are dimensionless. The order in which these constituents appear in any chain is biologically significant, and we encode this information in finite ordered sequences of arbitrary length L denoted $S\{L\} = (s_1, s_2, \ldots, s_L)$, whose elements s_j, $j = 1, \ldots L$ label the building blocks forming the chains, in the order indicated in the sequences. Each element s_j in the sequence $S\{L\}$ is an integer in the set $\{1, \ldots, n\}$ which refers to the type of building block occupying position j in the chain. There are therefore n^L sequences of length L if the model allows n types of building blocks. For instance, the sequence $S\{5\} = (1, 4, 3, 1, 3)$ in a model with, say, $n = 4$ types of building blocks (amino acids or codons), corresponds to a polymer chain of length 5 whose first component is a type 1 building block, the second component is a type 4 and so on. Given a sequence $S\{L\}$, we introduce subsequences $S\{L, j\} = (s_1, s_2, \ldots s_j)$ (resp. $S\widehat{\{L, j\}} = (s_{L-j+1}, s_{L-j+2}, \ldots s_L)$), $j = 1, \ldots L$, whose elements are the j leftmost (resp. rightmost) elements of $S\{L\}$. In particular, $S\{L, L\} \equiv S\widehat{\{L, L\}} \equiv S\{L\}$, $S\{L, 1\} = s_1$ and $S\widehat{\{L, 1\}} = s_L$. We write

$$S\{L\} = (S\{L, L-\ell\}, S\widehat{\{L, \ell\}}), \qquad 0 < \ell < L.$$

In what follows we sometimes refer to families of polymer chains differing only by their length and obtained by removing some rightmost building blocks from a chain of maximum length $L_{\max}$. Denoting by Π^S_ℓ a polymer chain of length ℓ and sequence $S\{\ell\}$ or subsequence $S\{L, \ell\}$, both having ℓ elements with $L > \ell$, the family of polymer chains obtained from a chain of maximal length $L_{\max}$ and sequence $S\{L_{\max}\}$ is given by $\{\Pi^S_\ell\}_{\ell=1,2,\ldots L_{\max}}$.

In the specific case of XNA/polynucleotide chains entering our model, we use $\Pi = R$ and the sequences are generically labeled as $\alpha\{\ell\}$. Their elements correspond to types of codons, and the complementary codon sequences in the sense of nucleic acids complementarity are $\bar{\alpha}\{\ell\}$. Therefore, a large class of XNA strands of length ℓ and sequence $\alpha\{\ell\}$ are denoted by R^α_ℓ, and in particular, $R^{\alpha_1}_1$ is a codon of type α_1. Besides the generic sequences $\alpha\{\ell\}$ introduced above, a sequence denoted $\pi\{L_{\max}\}$, together with its subsequences $\pi\{L_{\max}, \ell\}$ and $\pi\widehat{\{L_{\max}, \ell\}}$ for $\ell = 1, \ldots L_{\max}$ play a specific role: they correspond to polynucleotide chains that template the polymerization of a family of primordial peptide polymerases (p-Pol) through a process described in the next subsection, see also figure 3. Using $\Pi = P$ to denote polypeptide chains, this family of polymerases derived from $P_{L_{\max}}$ of maximal length $L_{\max}$, is $\{P^\pi_\ell\}_{\ell=2,\ldots,L_{\max}}$. These polymerases are such that $P^\pi_\ell = P^\pi_{\ell-1} + P^{\pi_\ell}_1$, with $P^{\pi_\ell}_1$ an amino acid π_ℓ. We use the notation P^π for a generic polymerase in the family. Alongside these polymerases, generic polypeptide chains of length ℓ and sequence $\alpha\{\ell\}$ are labeled as P^α_ℓ. Proteins of length 1, $P^{\alpha_1}_1$, are single amino acids of type α_1.

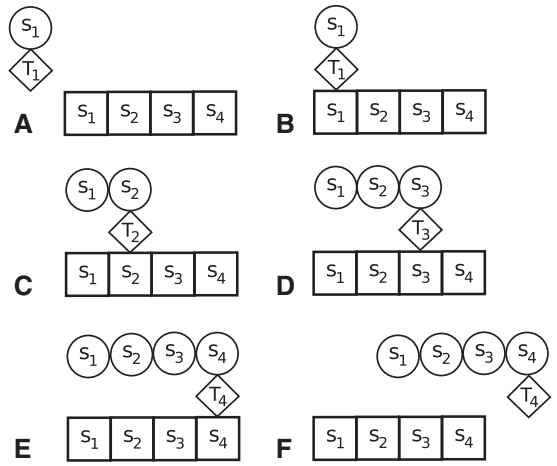

FIG. 3. Mechanism (B): Polypeptide polymerization in our model. The square boxes represent the codons of a polynucleotide chain (here, of length $L = 4$) and the circles represent amino acids. The p-tRNA molecules are labeled $T_1, \ldots, T_4$.

RNA–Protein Replication Scenario

The scenario relies on three types of mechanisms:

(A) The *spontaneous* polymerization of polynucleotide and polypeptide chains, assumed to occur at a very slow rate, and their depolymerization through being cleaved in two anywhere along the chains at a rate independent of where the cut occurs.

(B) The *nonspontaneous* polypeptide polymerization occurring through a polynucleotide chain R_L^S on which several p-tRNA molecules loaded with an amino acid dock and progressively build the polypeptide chain. More precisely, each codon of type s of the polynucleotide chain binds with a p-tRNA, itself linked to an amino acid of type s. Note that we assume the same number n of types of codons and amino acids. This leads to a chain of amino acids matching the codon sequence $S\{L\}$ of the polynucleotide chain. The process is illustrated in figure 3 for a polypeptide chain of length $L = 4$ and amino acid sequence $S\{4\} = (s_1, s_2, s_3, s_4)$.

(C) The duplication of a polynucleotide chain R_L^S, of length $L \geq \ell_{\pi\min}$, as a two-step process. In the first step, a polypeptide polymerase P^π, obtained by polymerization via mechanism (B) using a polynucleotide R_L^π, scans the polynucleotide chain R_L^S to generate its complementary polynucleotide chain $R_L^{\bar{S}}$. This is shown in figure 4. The resulting polynucleotide chain $R_L^{\bar{S}}$ is then used to generate a copy of the original polynucleotide chain R_L^S via the same mechanism (C).

The replicator crudely operates as follows:

- Mechanism (A) provides a small pool of polymer chains; among them, one finds short strands of XNA with dual function (p-mRNA and p-Rib)
- Mechanism (B) provides polypeptide chains, including the polymerases (p-Pol, called P^π here), by using the XNA produced through Mechanism (A) and Mechanism (C)

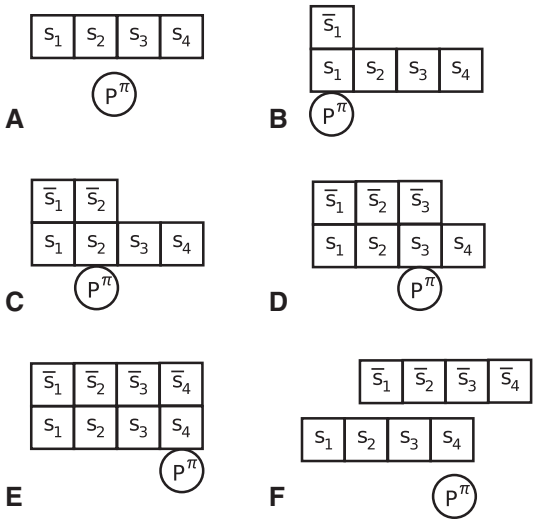

FIG. 4. First phase of Mechanism (C): Polymerization of the complementary polynucleotide chain $R_L^{\bar{S}}$ catalyzed by a primordial polymerase P^π.

- P^π are involved, through Mechanism (C), in the duplication of polynucleotides present in the environment, including the strands of XNA that participate in the very production of P^π

Reactions Driving the Replication and Physical Parameters

For simplicity, we consider the polymerization of polypeptide chains and the duplication of polynucleotide chains as single reactions where the reaction rates take into account all subprocesses as well as failure rates.

This leads to the following schematic reactions:

$$\text{Mechanism (A)}$$

$$R_L^S + R_1^{s_{L+1}} \rightarrow R_{L+1}^S, \tag{1}$$

$$R_L^S \rightarrow R_{L-\ell}^S + R_\ell^{\hat{S}} \qquad \ell = 1, \ldots, L-1, \tag{2}$$

$$P_L^S + P_1^{s_{L+1}} \rightarrow P_{L+1}^S, \tag{3}$$

$$P_L^S \rightarrow P_{L-\ell}^S + P_\ell^{\hat{S}} \qquad \ell = 1, \ldots, L-1. \tag{4}$$

$$\text{Mechanism (B)}$$

$$R_L^S + L \times TRP \rightarrow R_L^S + P_L^S. \tag{5}$$

$$\text{Mechanism (C)}$$

$$R_L^S + L \times R_1 \xrightarrow{P^\pi} R_L^S + R_L^{\bar{S}}, \tag{6}$$

where *TRP* denotes p-tRNA loaded with a single amino acid.

The parameters for these reactions are (see the Supplementary Material online for more details on the estimation of the parameter values):

- K_R^+: polymerization rate of polynucleotide chains (eq. 1); we have estimated the catalyzed XNA polymerization rate to be 4.2×10^{-7} mol^{-1} m^3 s^{-1}.

- K_R^-: depolymerization rate of polynucleotide chains (hydrolysis) (eq. 2); taken to be $8 \times 10^{-9} s^{-1}$.
- K_P^+: polymerization rate of polypeptide chains (eq. 3); we have estimated it to be 2.8×10^{-21} mol^{-1} m^3 s^{-1}.
- $K_{P,S,L}^-$: depolymerization rate of polypeptide chains of length L and sequence S (eq. 4); we have estimated it to be in the range $4 \times 10^{-11} s^{-1} - 5.1 \times 10^{-6} s^{-1}$.
- $k_{P,L}^+$: polymerization rate of a polypeptide of length L from the corresponding polynucleotide chain (eq. 5). It is reasonable to assume that $k_{P,L}^+ = k_{P,1}^+/L$ and we have estimated $k_{P,1}^+$ to be 0.1 mol^{-1} m^3 s^{-1}.
- Z: the rate at which a polymerase attaches to a polynucleotide chain (eq. 6) which we have estimated to be 10^6 mol^{-1} m^3 s^{-1}.
- h_R: the rate of attachment of a free polynucleotide to a polynucleotide chain attached to a p-Pol (eq. 6). We have estimated it to be 10^6 mol^{-1} m^3 s^{-1}.
- k_{step}: the rate at which a polymerase moves by one step on the polynucleotide (eq. 6). We have estimated it to be in the range $2 \times 10^{-2} s^{-1}$–$4 \times 10^{-5} s^{-1}$.

We now argue that the three parameters Z, h_R and k_{step} enter the dynamical system for the polymer concentrations in our model as two *physical* combinations denoted $\mathcal{K}(L)$ and $\mathcal{P}_b$ that we describe below.

First recall that we assume the existence of a pool of nucleotides, amino acids and p-tRNA. The amount of *free* nucleotides and amino acids is taken to be the difference between the total amount of these molecules and the total amount of the corresponding polymerized material, ensuring total conservation.

We denote the concentration of polypeptide and polynucleotide chains respectively by $P_L^\alpha, P_L^\pi, P_L^{\bar\pi}$ and $R_L^\alpha, R_L^\pi, R_L^{\bar\pi}$, all expressed in mol m^{-3}mol m^{-3}. In particular, P_1 and R_1 are the concentrations of each type of free amino acids and nucleotides respectively, and we assume, for simplicity, that all types of amino acids/codons are equally available.

We also assume that the amount of loaded p-tRNA, $C_{\text{p-tRNA}}$, remains proportional to the amount of free amino acids and that the concentration of p-tRNA is larger than P_1 so that most amino acids are loaded on a p-tRNA. With these conventions, one has

$$C_{\text{p-tRNA}} = k_t P_1 \qquad \text{with} \qquad k_t \approx 1. \tag{7}$$

Total Reaction Rate $\mathcal{K}(L)$ of Polynucleotide Polymerization

If a complex reaction is the result of one event at rate K, and m other, identical, events at rate k, the average time to complete the reaction is the sum of the average times for each event. Hence the reaction rate is given by

$$\tilde{\mathcal{K}}(K,k,m) = \left(\frac{1}{K} + \frac{m}{k}\right)^{-1} = \frac{Kk}{mK+k}. \tag{8}$$

One such complex reaction in our model is the polymerization of a polynucleotide chain of length L, say, from its complementary chain (second phase of Mechanism (C)). Polymerases are characterized by the polymerizing efficiency which, we assume, increases with ℓ, up to $L_{\max}$. The first step in polymerization requires a polymerase to attach itself to the template polynucleotide. This is only possible if the template polynucleotide has a minimum length, which we assume to be $\ell_{\pi\min}$. In the following, we assume that polymerases can polymerize polynucleotide chains of any length greater or equal to $\ell_{\pi\min}$. The corresponding reaction rate is given by $Z P_\ell^\pi$ for a polymerase of length $\ell \geq \ell_{\pi\min}$.

The free nucleotides must then attach themselves to the polynucleotide–polymerase complex and the polymerase must move one step along the polynucleotide. The rate for each of these L steps is

$$k_{R+} = \frac{k_{\text{step}} h_R R_1}{k_{\text{step}} + h_R R_1}, \tag{9}$$

and hence, the rate of polymerization for a polynucleotide of length L and polymerase of length ℓ is $\tilde{\mathcal{K}}(Z P_\ell^\pi, k_{R+}, L)$. However, it is assumed that polymerases of several lengths are available and therefore, the total rate is given by

$$\mathcal{K}(L) = \begin{cases} \sum_{\ell=\ell_{\pi\min}}^{L_{\max}} \tilde{\mathcal{K}}(Z P_\ell^\pi, k_{R+}, L)\, W_\ell, & L \geq \ell_{\pi\min} \\ 0 & L < \ell_{\pi\min}, \end{cases} \tag{10}$$

where it is understood that $\ell_{\pi\min}$ is the lower bound length for polymerase activity and W_ℓ is a quality factor given by

$$W_\ell = \begin{cases} \dfrac{\ell - \ell_{\pi\min} + 1}{\ell_{\pi\max} - \ell_{\pi\min} + 1} & \ell_{\pi\min} \leq \ell \leq \ell_{\pi\max} \\ 1 & \ell_{\pi\max} < \ell \leq L_{\max}. \end{cases} \tag{11}$$

Indeed, we expect long polymerases to be more efficient, so W_ℓ is taken to increase with ℓ in the range $\ell_{\pi\min} \leq \ell \leq \ell_{\pi\max}$, while polymerases of length $\ell > \ell_{\pi\max}$ have the same level of activity as those with length $\ell = \ell_{\pi\max}$, that is, $W_{\ell > \ell_{\max}} = 1$.

To avoid proliferation of parameters in our simulations, we have taken $\ell_{\pi\max} = L_{\max}$, where $L_{\max}$ is the maximal polynucleotide chain's length.

Binding Probability $\mathcal{P}_b$ of a Polynucleotide and a Polymerase of Length L

First note that it takes L times longer to synthesize a polypeptide chain of length L from its corresponding polynucleotide chain than it takes for one amino acid to bind itself to the polynucleotide. The rate is thus given by $k_{P,L}^+ P_1 = (k_{P,1}^+/L) P_1$.

We now offer some considerations on depolymerization. We assume that if a polymer Π_L^S depolymerizes, it does so by (potentially consecutive) cleavings. In the first step, Π_L^S can cleave in $L - 1$ different positions, resulting in two smaller chains L_1, L_2 with $L = L_1 + L_2$ and $1 \leq L_{1,2} \leq L - 1$. This is the origin of the factor $(L - 1)$ in the terms describing the depolymerization of polymer chains in the dynamical systems equations presented in the next subsection.

The concentration variations resulting from such depolymerizations must be carefully evaluated. A polymer Π_L^S of length L and sequence S, where S stands for any of α, π or $\bar\pi$, can be obtained by cleaving a polymer $\Pi_\ell^{\tilde S}$ of length

$\ell > L$ and sequence $\tilde{S} = (S, T)$ where T is a sequence of length $\ell - L$. Similarly it can be obtained by cleaving $\Pi_\ell^{\tilde{S}'}$ of sequence $\tilde{S}' = (T', S)$ where T' is also of length $\ell - L$. If the rate of cleaving, K_Π^-, is assumed to be independent of the polymer length, and since there are $n^{\ell-L}$ different sequences T and T', where n is the number of amino acid or codon *types*, the rate of concentration variation of polymers of length L resulting from the depolymerization of longer polymers is

$$\sum_{\ell=L+1}^{L_{\max}} n^{\ell-L} K_\Pi^- \Pi_\ell^{\tilde{S}} + \sum_{\ell=L+1}^{L_{\max}} n^{\ell-L} K_\Pi^- \Pi_\ell^{\tilde{S}'}. \quad (12)$$

Recall that we use the same notation for the concentration of a polymer of sequence S and length L and the polymer itself, namely Π_L^S, and Π is supposed to be set to $\Pi = P$ or $\Pi = R$ in our model. As already stressed, we assume polymers have at most length $L_{\max}$. Finally, when the concentrations $\Pi_L^{\tilde{S}}$ and $\Pi_L^{\tilde{S}'}$ are equal, (eq. 12) can be rewritten as

$$2 \sum_{\ell=L+1}^{L_{\max}} n^{\ell-L} K_\Pi^- \Pi_\ell^{\tilde{S}}. \quad (13)$$

The depolymerization of polymerase P_L^π requires special treatment. When P_L^π depolymerizes, it generates a polymerase P_ℓ^π with $\ell < L$. On the other hand, any P_L^π can be obtained through depolymerization of one of $2n$ types of polymers of length $L+1$, one of which being P_{L+1}^π and the remaining $2n-1$ being of type P_{L+1}^α with $\alpha\{L+1\} = (\pi\{L\}, \alpha_{L+1})$, $\alpha_{L+1} \neq \pi_{L+1}$, or $\alpha\{L+1\} = (\alpha_1, \pi\{L\})$ with α_1 any of the n types of amino acids. More generally, they can be obtained from $P_{L+\ell'}^\pi$ and $2n^{\ell'} - 1$ polymers of type $P_{L+\ell'}^\alpha$ where $\ell' \geq 1$ and $\alpha\{L+\ell'\} = (\pi\{L\}, \alpha_{L+1}, \ldots \alpha_{L+\ell'})$ with $\alpha_j \neq \pi_j, j = L+1, \ldots L+\ell'$, or $\alpha\{L+\ell'\} = (\alpha_1, \ldots \alpha_{\ell'}, \pi\{L\})$ for any type $\alpha_j, j = 1, \ldots \ell'$. The same is true for the corresponding polynucleotide chains.

When the polymerase is bound to a polynucleotide, it becomes more stable either through induced folding of a (partially) unfolded sequence, or through the inaccessibility of bound portions, or both. We thus define $F_\pi(\ell)$ as the depolymerization reduction coefficient for the bound polymerase of length ℓ, with that reduction coefficient being 1 when no depolymerization occurs at all. We estimate it to be

$$F_\pi(\ell) = \begin{cases} 1 - e^{-\frac{\ell - \ell_{\pi\min} + 1}{\lambda}} & \ell \geq \ell_{\pi\min} \\ 0 & \ell < \ell_{\pi\min} \end{cases}, \quad (14)$$

with $\lambda > 0$ a parameter controlling how much of the polymerase is stabilized. The term $(\ell - \ell_{\pi\min} + 1)/\lambda$ can be interpreted as a Boltzmann factor with a free energy expressed in units of k_BT. The hydrogen bond binding energy between RNA and a polypeptide is ~16 kJ/mol [Dixit et al. 2000], so assuming that the number of such hydrogen bonds between the polymerase and the polynucleotide is $\ell - \ell_{\pi\min} + 1$, one has $\lambda \approx 0.15$.

The binding rate of a polymerase to a polynucleotide R_M^α of length M and sequence α is $k_{b,M} = Z R_M^\alpha n^M$ where n^M is the total number of polynucleotides of length M. The probability that a polymerase of length L binds to a polynucleotide of length M is therefore given by

$$\tilde{\mathcal{P}}_{b,M} = \frac{k_{b,M}}{\sum_{m=2}^{L_{\max}} k_{b,m}}. \quad (15)$$

The total time the polymerase remains bound to a polynucleotide of length M is estimated to be M/k_{R+}. Therefore the probability $\mathcal{P}_b$ for a polymerase to be bound is given by the average binding time divided by the sum of the average binding time and the average time needed to bind:

$$\mathcal{P}_b = \frac{\sum_{M=2}^{L_{\max}} (M/k_{R+}) \tilde{\mathcal{P}}_{b,M}}{\sum_{M=2}^{L_{\max}} ((M/k_{R+}) \tilde{\mathcal{P}}_{b,M}) + 1/\sum_{m=2}^{L_{\max}} k_{b,m}}. \quad (16)$$

As a result the polymerase depolymerization rate will be

$$\begin{aligned} K_{P,\alpha,L}^- &= K_P^-, \\ K_{P,\bar{\pi},L}^- &= K_P^-, \\ K_{P,\pi,L}^- &= K_P^- (1 - \mathcal{P}_b F_\pi(L)). \end{aligned} \quad (17)$$

Equations

For any chain of length ℓ, our model considers the concentrations of polynucleotides and polypeptides corresponding to the polymerase sequence π, its complementary sequence $\bar{\pi}$ and the generic sequences α. We assume that the concentrations of polynucleotides and polypeptides of a specific length, bar the polymerase and its complementary sequence, are identical. For the chains that share the first ℓ elements of their sequence with those of the polymerase (or its complementary chain), and differ in all other elements, this is only an approximation, but it is nevertheless justified, as the concentrations of these polymers only differ slightly from those of polymers with sequences of type α, and their contribution to the variation of the polymerase concentration is expected to be small.

The variations in polymer concentrations as time evolves are governed in our model by a system of ordinary differential equations. In the equations, L is the length of the polymer chains, spanning all values in the range $1 < L \leq L_{\max}$ where $L_{\max}$ is the maximal length of polypeptide and polynucleotide chains. We thus have a system of $6 \times (L_{\max} - 1)$ equations. We recall that n is the number of codon types, assumed to be equal to the number of amino acid types.

$$\begin{aligned} \frac{dR_L^\pi}{dt} &= K_R^+ R_1 R_{L-1}^\pi - n K_R^+ R_1 R_L^\pi \\ &\quad + \sum_{\ell=L+1}^{L_{\max}} [K_R^- R_\ell^\pi + (2n^{\ell-L} - 1) K_R^- R_\ell^\alpha] \\ &\quad - (L-1) K_R^- R_L^\pi + \mathcal{K}(L) R_L^{\bar{\pi}}, \end{aligned}$$

$$\begin{aligned} \frac{dR_L^{\bar{\pi}}}{dt} &= K_R^+ R_1 R_{L-1}^{\bar{\pi}} - n K_R^+ R_1 R_L^{\bar{\pi}} \\ &\quad + \sum_{\ell=L+1}^{L_{\max}} [K_R^- R_\ell^{\bar{\pi}} + (2n^{\ell-L} - 1) K_R^- R_\ell^\alpha] \\ &\quad - (L-1) K_R^- R_L^{\bar{\pi}} + \mathcal{K}(L) R_L^\pi, \end{aligned}$$

$$\frac{dR_L^\alpha}{dt} = K_R^+ R_1 R_{L-1}^\alpha - nK_R^+ R_1 R_L^\alpha + 2\sum_{\ell=L+1}^{L_{\max}} n^{\ell-L} K_R^- R_\ell^\alpha - (L-1)K_R^- R_L^\alpha + \mathcal{K}(L) R_L^\alpha,$$

$$\frac{dP_L^\pi}{dt} = K_P^+ P_1 P_{L-1}^\pi - nK_P^+ P_1 P_L^\pi + \sum_{\ell=L+1}^{L_{\max}} [K_P^- (1-\mathcal{P}_b F_\pi(L)) P_\ell^\pi + (2n^{\ell-L}-1)K_P^- P_\ell^\alpha] - (L-1)K_P^- (1-\mathcal{P}_b F_\pi(L) P_L^\pi + k_{P,L}^+ P_1 R_L^\pi,$$

$$\frac{dP_L^{\bar\pi}}{dt} = K_P^+ P_1 P_{L-1}^{\bar\pi} - nK_P^+ P_1 P_L^{\bar\pi} + \sum_{\ell=L+1}^{L_{\max}} [K_P^- P_\ell^{\bar\pi} + (2n^{\ell-L}-1)K_P^- P_\ell^\alpha] - (L-1)K_P^- P_L^{\bar\pi} + k_{P,L}^+ P_1 R_L^{\bar\pi},$$

$$\frac{dP_L^\alpha}{dt} = K_P^+ P_1 P_{L-1}^\alpha - nK_P^+ P_1 P_L^\alpha + 2\sum_{\ell=L+1}^{L_{\max}} n^{\ell-L} K_P^- P_\ell^\alpha - (L-1)K_P^- P_L^\alpha + k_{P,L}^+ P_1 R_L^\alpha. \quad (18)$$

Alongside the seven physical parameters $\{K_R^\pm, K_P^\pm, h_{P,L}^+, \mathcal{K}(L), \mathcal{P}_b\}$ appearing in the differential equations above, we need to consider two parameters yielding the "initial" concentrations of amino acid and nucleotide inside the system, namely $\rho_p \equiv P_1(t=0)$ and $\rho_r \equiv R_1(t=0)$. In the absence of actual data for these quantities, we explore a range of realistic values in the analysis of our model. The concentration of free amino acids and nucleotides at any one time is then given by $P_1(t) = \rho_p - \sum_{L=2}^{L_{\max}}[(n^L-2)P_L^\alpha(t) + P_L^\pi(t) + P_L^{\bar\pi}(t)]$ and $R_1(t) = \rho_r - \sum_{L=2}^{L_{\max}}[(n^L-2)R_L^\alpha(t) + R_L^\pi(t) + R_L^{\bar\pi}(t)]$ respectively, with $P_L^S(0) = R_L^S(0) = 0$ for any value of L in the range $2 \le L \le L_{\max}$ and sequence $S = \alpha, \pi, \bar\pi$.

Results

The system of equations (eq. 18) is nonlinear and too complex to solve analytically. We therefore analyze it numerically, starting from a system made entirely of free nucleotides, amino acids, as well as charged p-tRNA, and letting the system evolve until it settles into a steady configuration.

The main quantities of interest are the relative concentrations of the polymerase (ρ_π) and of the α peptide chains (ρ_α). We have

$$\rho_\pi = \sum_{\ell=\ell_{\pi\min}}^{L_{\max}} P_\ell^\pi \quad \text{and} \quad \rho_\alpha = \sum_{\ell=\ell_{\pi\min}}^{L_{\max}} P_\ell^\alpha, \quad (19)$$

and evaluate the ratios

$$Q_1 = \frac{\rho_\pi}{\rho_\alpha} \quad \text{and} \quad Q_{2,\ell} = \frac{P_\ell^\pi}{P_\ell^\alpha}, \quad (20)$$

while monitoring the evolution of each quantity over time. Q_1 corresponds to the relative amount of polymerase of any length compared with other proteins (for a specific arbitrary sequence α), while $Q_{2,\ell}$ corresponds to the relative amount of polymerase of length ℓ compared with an arbitrary protein of length ℓ. Unit ratios indicate that the polymerase has not been selected at all, whereas large values of Q_1 or $Q_{2,\ell}$ on the other hand indicate a good selection of the polymerase.

The complexity of the system (eq. 18) also lies in the number of free parameters it involves. A systematic analysis of the high-dimensional parameter space is beyond the scope of this article, and we therefore concentrate on the analysis and description of results for a selection of parameter values that highlight potentially interesting behaviors of our model.

Recall that our model assumes that the number n of different amino acids is equal to the number of codon types, and throughout our numerical work we have set $n = 4$. Note that the word "codon" here is used by extension. Indeed, there are four different nucleic acids in our model and the "biological" codons are made of two nucleic acids, bringing their number to sixteen. However, they split into four groups of four, each of which encoding one of the four amino acids. From a mathematical modeling point of view, this is completely equivalent. It is well accepted that early proteins were produced using a reduced set of amino acids (Angyan et al. 2014). The exact identity and number is unclear though experimental work has shown that protein domains can be made using predominantly five amino acids (Riddle et al. 1997) whereas the helices of a four-alpha helix bundle were made using only four amino acids (Regan and DeGrado 1988). We have used mostly $\ell_{\pi\min} = 7$ and $\ell_{\pi\max} = L_{\max} = 10$, but have investigated other values as well (see the Supplementary Material online).

While these figures are somewhat arbitrary, an $\ell_{\pi\min}$ of 7 was chosen on the assumption that the functional p-Pol would have some forms of stable structural motif and this number corresponds to the typical minimum number of amino acids required to produce a stable, folded alpha helix structure (Manning et al. 1988). The choice of $L_{\max} = 10$ is based on the fact that while the polymer peptide chains could be significantly longer, they would need correspondingly long polynucleotide sequences to encode them, which becomes increasingly unlikely as lengths increase. Furthermore, we expected polymers of length 10 to have very low concentrations, a hypothesis confirmed by our simulations. We have nevertheless investigated larger values of $L_{\max}$ as well, and found little difference, as outlined below.

In a first step, guided by data on parameter values gleaned from the literature and gathered in the Supplementary Material online, we set

$$\begin{aligned} K_R^+ &= 4.2 \times 10^{-7}\,\text{mol}^{-1}\,\text{m}^3\,\text{s}^{-1},\\ K_R^- &= 8 \times 10^{-9}\text{s}^{-1},\\ K_P^+ &= 2.8 \times 10^{-21}\,\text{mol}^{-1}\,\text{m}^3\,\text{s}^{-1},\\ K_P^- &= 4 \times 10^{-11}\text{s}^{-1}\\ k_{P,1}^+ &= 0.1\,\text{mol}^{-1}\,\text{m}^3\,\text{s}^{-1},\\ h_R &= 10^6\,\text{mol}^{-1}\,\text{m}^3\,\text{s}^{-1},\\ Z &= 10^6\,\text{mol}^{-1}\,\text{m}^3\,\text{s}^{-1},\\ \lambda &= 0.15,\\ k_{\text{step}} &= 4 \times 10^{-5}\text{s}^{-1}. \end{aligned} \quad (21)$$

We let the system evolve under a variety of initial concentrations of free amino acids and nucleotides, ρ_p and ρ_r, in the range $10^{-5} - 0.1\,\mathrm{mol\,m^{-3}}$, and with all polymer concentrations set to 0. We monitored the concentration of all polymers, in particular the concentration of polymerase ρ_π and its ratio to the concentration of α polypeptide chains, Q_1. In most cases we found that the nucleotides polymerized spontaneously (Mechanism (A)) in small amount and this led, indirectly, to the polymerization of the polypeptides, including the polymerases (Mechanism (C)). The polymerases then induced further polymerization of the polynucleotides (Mechanism (B)) and the system slowly equilibrated.

The end result was an excess of polymerase of all lengths compared with α polypeptide chains with $Q_1 = 786$ for all initial concentrations $\rho_p = \rho_r \geq 0.001\,\mathrm{mol\,m^{-3}}$ (fig. 5). Moreover the total amount of polymerase reached, for initial concentration of free amino acids ρ_p, was a concentration of $\sim 4 \times 10^{-4} \times \rho_p$ (as illustrated by the bottom two rows in table 1). The concentration of polymerase of length 10, on the other hand, was very small $P_{10}^{\pi} = 6.3 \times 10^{-14}\mathrm{mol\,m^{-3}}$ for but $Q_{2,10} = 5.9 \times 10^{18}$ was very large, effectively showing that the only polypeptide chain of length $L_{\max} = 10$ was the polymerase.

We found hardly any polymerization of the polymerase when $\rho_p = \rho_r = 0.0009\,\mathrm{mol\,m^{-3}}$, with $\rho_\pi \approx 1.4 \times 10^{-14}\ \mathrm{mol\,m^{-3}}$ and $Q_1 = 12.4$, whereas with $\rho_p = \rho_r = 0.001\,\mathrm{mol\,m^{-3}}$, we obtained $\rho_\pi \approx 3.9 \times 10^{-7}\,\mathrm{mol\,m^{-3}}$ and $Q_1 = 786$ (fig. 5*a*). This highlights a very sharp transition at a *critical concentration* $\rho_{p,c}$ above which polymerases are generated. We summarize the data in table 1.

We then fixed the initial concentration ρ_p to four different values and varied ρ_r to identify the critical initial concentration of nucleotides necessary for the production of polymerases. The results in table 2 show that the critical concentration $\rho_{r,c}$ is nearly constant and of the order of $10^{-3}\,\mathrm{mol\,m^{-3}}$ for a very wide range of amino acid initial concentrations.

Many of the parameters we have used were estimated or measured in conditions which, in all likelihood, were not identical to the ones existing when the polymerization we are modeling occurred. In a second step, we departed from the set of values (eq. 21) and found that in all cases investigated, varying these parameters modified the critical concentrations of $\rho_{r,c}$ and $\rho_{p,c}$, but did not affect significantly the value of Q_1 while $Q_{2,10}$ remained extremely large.

More specifically, taking $K_P^- = 5.1 \times 10^{-6}\mathrm{s^{-1}}$ marginally increased the critical concentration to $\rho_{r,c} = \rho_{p,c} = 0.0011\,\mathrm{mol\,m^{-3}}$. Similarly, taking $k_{\mathrm{step}} = 0.02\mathrm{s^{-1}}$ increased slightly the critical concentrations: $\rho_{r,c} = \rho_{p,c} = 0.0017\,\mathrm{mol\,m^{-3}}$. On the other hand, taking $Z = 10^8\mathrm{mol^{-1}\,m^3\,s^{-1}}$ lead to a decrease of the critical concentrations: $\rho_{r,c} = \rho_{p,c} = 0.0005\,\mathrm{mol\,m^{-3}}$. Varying h_R to values as small as $1\,\mathrm{mol^{-1}\,m^3\,s^{-1}}$ did not change the critical concentrations.

In our model, we have considered the concentrations of free amino acids ($\rho_p \equiv P_1$) and charged p-tRNA to be identical: $k_t \approx 1$ (see eq. 7). To consider other values of k_t we only need to multiply the polymerization rate of a peptide ($k_{P,1}^+$) by k_t as it is p-tRNAs that bind to XNA chains, not free amino acids. We have considered a large range of values for $k_{P,1}^+$ and found that for $k_{P,1}^+ = 10^{-5}\,\mathrm{mol^{-1}\,m^3\,s^{-1}}$, the critical concentrations had not changed significantly while for $10^{-8}\,\mathrm{mol^{-1}\,m^3\,s^{-1}}$, they increased to $\rho_{r,c} = \rho_{p,c} = 0.002\,\mathrm{mol\,m^{-3}}$. This shows that taking much smaller values of k_t has a very small impact on our results and that having a concentration of charged p-tRNA much smaller than that of free amino acids would only increase marginally the critical concentrations we have obtained using our original assumption.

The parameters on which the model is the most sensitive are K_R^+ and K_R^-. We found that for $K_R^+ = 4 \times 10^{-8}\,\mathrm{mol^{-1}\,m^3\,s^{-1}}$, $\rho_{r,c} = \rho_{p,c} = 0.007\,\mathrm{mol\,m^{-3}}$ and for $K_R^+ = 4 \times 10^{-9}\,\mathrm{mol^{-1}\,m^3\,s^{-1}}$, $\rho_{r,c} = \rho_{p,c} = 0.05\,\mathrm{mol\,m^{-3}}$. Similarly, for $K_R^- = 10^{-7}\,\mathrm{s^{-1}}$ we found that $\rho_{r,c} = \rho_{p,c} = 0.01\,\mathrm{mol\,m^{-3}}$ and

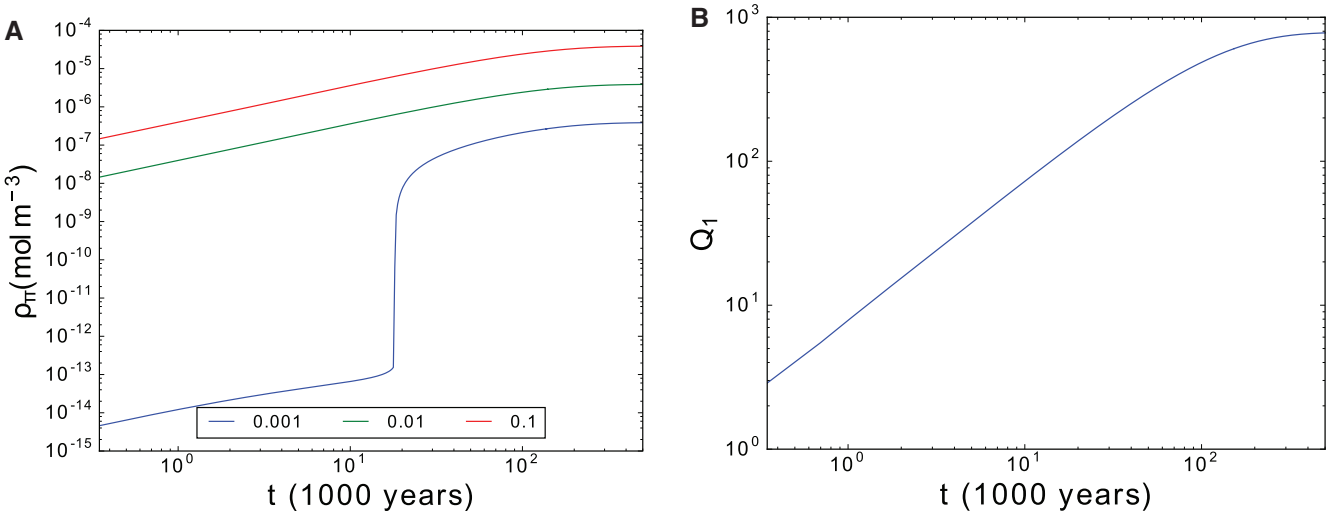

FIG. 5. (*a*) Time evolution of the polymerase for initial concentration $\rho_r = \rho_p = 0.001$, 0.01, and $0.1\,\mathrm{mol\,m^{-3}}$. (*b*) Q_1 for initial concentration $\rho_r = \rho_r = 0.01\,\mathrm{molm^{-3}}$. Parameter values: $K_P^- = 4 \times 10^{-11}\mathrm{s^{-1}}$, $Z = 10^6\,\mathrm{mol^{-1}\,m^3\,s^{-1}}$, $\lambda = 0.15$.

Table 1. Effect of Initial Concentrations on Polymerase Production.

ρ_p (mol m^{-3})	1. ρ_r (mol m^{-3})	2. ρ_π (mol m^{-3})	3. Q_1	4. Polymerase Production
2×10^{-4}	2×10^{-4}	$2.8\ 10^{-19}$	1.0008	Insignificant
9×10^{-4}	9×10^{-4}	1.410^{-14}	12.4	Insignificant
10^{-3}	10^{-3}	3.9×10^{-7}	786	Yes
10^{-1}	10^{-1}	3.9×10^{-5}	786	Yes

Table 2. Effect of Initial Peptide Concentration on Critical Concentration.

ρ_p (mol m^{-3})	$\rho_{r,c}$ (mol m^{-3})
10^{-4}	2×10^{-3}
10^{-3}	10^{-3}
10^{-2}	8×10^{-4}
10^{-1}	7×10^{-4}

Table 3. Effect of Initial Concentration of Free Nucleotides on Time for Production of Polymerase.

$\rho_r = \rho_p$ (mol m^{-3})	τ_{pol} (*years*)
0.001	18,000
0.002	254
0.005	12.7
0.01	2.2

for $K_R^- = 10^{-6}\,s^{-1}$ that $\rho_{r,c} = \rho_{p,c} \approx 0.05\,mol\,m^{-3}$. This shows that the spontaneous polymerization of polynucleotide is essential to reach a minimum concentration of polynucleotides to kick start the whole catalysis process and that the stability of the polynucleotides plays an important role.

To investigated this, we have run simulations with $K_R^+ = 4 \times 10^{-8}\,mol^{-1}\,m^3\,s^{-1}$ for a fixed duration, τ_{pol}, after which K_R^+ was set to 0. We found that if τ_{pol} was long enough, the polymerization of polypeptide and polynucleotide chains was identical to the one obtained whereas K_R^+ was not modified. When τ_{pol} was too short, on the other hand, one was only left with short polypeptide and polynucleotide chains in an equilibrium controlled by the spontaneous polymerization and depolymerization parameters. The minimum value for τ_{pol} depends on the concentrations ρ_r and ρ_p and the results are given in table 3.

This shows that while K_R^+ is an important parameter in the process, what matters are to have a spontaneous generation of polynucleotides at the onset (Mechanism (A)). This then leads to the production of polypeptides, including polymerase (Mechanism (C)) and, once the concentration of polymerase is large enough, the catalyzed production of polynucleotides (Mechanism (B)) dominates the spontaneous polymerization.

We have also varied K_R^- once the system had settled and we found that for $\rho_r = \rho_p = 0.01\,mol\,m^{-3}$, K_R^- could be increased up to $6 \times 10^{-7} s^{-1}$ while still keeping a large amount of polymerase. Above that value, the polynucleotides are too unstable and one ends up again with mostly short polymer chains and $Q_1 \approx 1$.

Table 4. Effect of Various Parameters on Initial Critical Concentrations.

Modified Parameter	$\rho_{r,c} = \rho_{p,c}$ (mol m^{-3})
$K_p^- = 5.1 \times 10^{-6}\,s^{-1}$	1.1×10^{-3}
$k_{step} = 2 \times 10^{-2}\,s^{-1}$	1.7×10^{-3}
$Z = 10^8\,mol^{-1}\,m^{-3}s^{-1}$	5×10^{-4}
$h_R = 1\,mol^{-1}\,m^{-3}s^{-1}$	10^{-3}
$k_{p,1}^+ = 10^{-5}\,mol^{-1}\,m^3\,s^{-1}$	10^{-3}
$k_{p,1}^+ = 10^{-8}\,mol^{-1}\,m^3\,s^{-1}$	2×10^{-3}
$L_{max} = 15$	1.1×10^{-3}
$L_{\pi min} = 6$	4×10^{-4}
$L_{\pi min} = 5$	2×10^{-4}
$L_{\pi min} = 4$	2×10^{-5}
$K_R^+ = 4 \times 10^{-8}\,mol^{-1}\,m^3\,s^{-1}$	7×10^{-3}
$K_R^+ = 4 \times 10^{-9}\,mol^{-1}\,m^3\,s^{-1}$	5×10^{-2}
$K_R^- = 10^{-7}\,s^{-1}$	10^{-2}
$K_R^- = 10^{-6}\,s^{-1}$	0.19

We have also considered values of $L_{max} > 10$ and found that the main difference is a slight increase of the critical concentrations. For example, for $L_{max} = 11, 12$ and 15, $\rho_{r,c} = \rho_{p,c}$ are respectively equal to $0.001, 0.0011$, and $0.0011\,mol\,m^{-3}$. At given concentrations Q_1 and ρ_π remain unchanged but $P^\pi_{L_{max}}$ deceases approximately by a factor of 40 each time L_{max} is increased by 1 unit.

We have also taken $L_{\pi min} = 4$, 5, and 6 and found that the critical concentrations were respectively $2 \times 10^{-5}, 2 \times 10^{-4}$, and $4 \times 10^{-4}\,mol\,m^{-3}$, whereas ρ_π took the values of $\sim 0.012, 2.6 \times 10^{-3}$, and $3 \times 10^{-4}\,mol\,m^{-3}$. Q_1 on the other hand remained constant.

A summary of the parameter values investigated outside the set (eq. 21) and the corresponding critical concentrations are given in table 4. Only one parameter was changed at a time (see the Supplementary Material online).

Discussion

We describe a theoretical nucleopeptidic reciprocal replicator comprising a polynucleotide that templates the assembly of small p-tRNA adapter molecules, most likely having mixed backbone architectures. These spontaneously arising p-tRNAs would have been bound to various classes of amino acids (possibly via weak stereochemical specificity), and a simple increase in local concentration mediated by binding to the p-Rib (in its most primitive version nothing much more than a mixed backbone architecture p-mRNA) could have driven polypeptide polymerization. Once a template arose that coded for a peptide able to catalyze phosphodiester bond formation, this p-Rib could have templated assembly of its own complementary strand (and vice versa) and the self-replication cycle would have been complete (see fig. 6 for a summary).

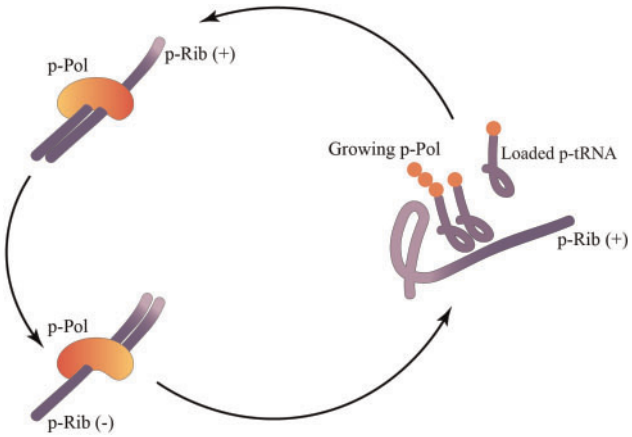

Fig. 6. The nucleopeptide Initial Darwinian Ancestor. In this cartoon model, a short strand of XNA has the functionality of both a primordial p-mRNA and a p-Rib. Primitive XNA molecules loaded with amino acids (p-tRNA) bind to the p-Rib via codon–anticodon pairing. This allows adjacent amino acids to undergo peptide bond formation and a short peptide chain is produced. A certain peptide sequence is able to act as a primordial XNA-dependent XNA polymerase (p-Pol) able to copy both + and – p-Rib strands to eventually produce a copy of the p-Rib(+) strand.

Starting from a single peptide and single polynucleotide, the IDA would quickly have become a distribution of related sequences of peptides and XNAs. We can imagine that over time, different p-Ribs encoding different peptides with additional functionalities could have appeared as the system evolved and that these p-Ribs may have subsequently fused together into larger molecules.

By imagining the IDA swiftly becoming a pool of molecules where variety within the "species" is maintained by the poor copying fidelity of a statistical operational code, should any mutation that stops replication arise, the other molecules in the pool would still function, ensuring continuity of the whole. Indeed this could have provided a selective pressure for superior replicators. While our model does not directly consider less than perfect copying fidelity, it is not expected to have a major effect on our conclusions as copies with decreased performance would not be maintained as a significant proportion of the population and copies with increased performance would simply take over the role of main replicator.

The primordial operational code may only have required two bases per p-tRNA to deliver statistical proteins, while the catalytic requirements of the p-Pol are loose enough that a seven-residue peptide is a plausible lower length limit. This reduces the minimum length of the posited spontaneously arising p-Rib to just 14 nucleotides (assuming no spaces between codons). This is an optimistic length estimate, but given the available time and with molecular co-evolution, inorganic catalysts and geological PCR, considerably longer molecules may have been possible (Baaske et al. 2007; Fishkis 2011). These p-Pols would act on p-Ribs and the crucial abiogenesis step would be the emergence of a 14-mer XNA that, in the context of the primordial operational code, happened to code for a peptide able to bind XNA and catalyze phosphodiester bond formation of base-paired nucleotides. Although the concentrations of various components are not known with certainty this does not seem an unreasonable proposition particularly given that functional peptides are known to occur in random sequences with surprising frequency (Keefe and Szostak 2001).

Our mathematical model showed that the most important parameters, apart from the concentration of loaded p-tRNA and polynucleotides, are the spontaneous polymerization and depolymerization of polynucleotides. It also shows that polynucleotides are first polymerized spontaneously and that these initial polynucleotides catalyze the production of the first polypeptides, including the polymerase. These polymerases can then generate further polynucleotides through catalysis. The stability gained by polymerases while being bound to polynucleotides ultimately leads to an increase of their relative concentration compared with the other polypeptides.

Overall, the hypothesis explains the coupling of polynucleotide and polypeptide polymerization, the operational code and mutations in the p-Pol sequence that could eventually result in increased specificities leading to primitive DNA polymerases and RNA polymerases. No extraordinary exchanges of function are required and each molecule is functionally similar to its present-day analogue. Like all new abiogenesis theories, this IDA requires in vitro confirmation; in particular, the steps required for the primordial operational code to arise ab initio warrant close attention.

The idea that the ancestral replicator may have consisted of both nucleic acid and peptide components (the "nucleopeptide world") is in itself not new, but compared with the RNA world, has been somewhat neglected. We argue that molecular co-evolution of polynucleotides and peptides seems likely and cross-catalysis is known to be possible, for example in vitro selection experiments delivered RNA with peptidyl transferase activity after just nine rounds of a single selection experiments (Zhang and Cech 1997; Fishkis 2011). Inversely, Levy and Ellington produced a 17-residue peptide that ligates a 35 base RNA (Levy and Ellington 2003).

Nucleopeptide world research is relatively sparse, the data collected so far hint that cross-catalysis may be more efficient than autocatalysis by either peptides or nucleic acids. A self-replicating primordial system wherein RNA encoding for protein was replicated by a primordial RNA-dependent RNA polymerase which carried out the role of a replicative agent rather than as a transcriber of genes has previously been suggested (Leipe et al. 1999), although in this case no further development of the concept to produce a self-contained replicating system was pursued. The merits of a "two polymerase" system where RNA catalyses peptide polymerization and vice versa were succinctly explained by Kunin (2000), although possible mechanisms and validity were not considered in detail. The possibility of a two polymerase system is also mentioned by van der Gulik and Speijer as part of a wider review of the co-evolution of peptides and RNA (van der Gulik and Speijer 2015) but without a mathematical model.

Other origins of life hypotheses propose that the initial self-replicator did not consist of polynucleotides and/or peptides but was originally composed of different materials, most

famously clay crystals (Cairns-Smith 1982). Such hypotheses are of interest but were not considered in this work as the IDA presented here does not require genetic takeover of one replication system by another and can be achieved using building blocks likely to have been present on the early earth and so appears more parsimonious. Our IDA hypothesis has tried to set out more rigorously the possible steps and processes whereby a nucleopeptide IDA could have arisen and could be tested experimentally.

Future experimental work that would support the nucleopeptide theory would be to provide evidence that the stereochemical hypothesis applies to the earliest occurring amino acids including those likely to have composed the active site of the p-Pol. Currently codon/anticodon binding to a number of amino acids has been shown (Yarus et al. 2005) but is absent for the four earliest amino acids (Wolf and Koonin 2007). This may be due to their small sizes though even here possible solutions have been proposed (Tamura 2015).

It is important to note that we do not propose that the RNA world did not or could not exist, nor does this work necessarily suggest that a self-replicating RNA polymerase did not exist (although our results suggest it to be unlikely), but rather that such a molecule did not directly lead to current living systems. Indeed the crucial role of RNA (more correctly, XNA) in our model is highlighted by the importance of K_R^+, the rate of polymerization of polynucleotide chains. We also do not dismiss any roles for ribozymes—for example it could well be that ribozymes were responsible for aminoacylation reactions (although this would inevitably raise the question of how such ribozymes were themselves replicated). Similarly (and with similar provisos), peptides alone could also have carried out supporting roles such as stabilizing long XNA sequences or catalyzing aminoacylation reactions. At its core however, we suggest that the ancestral replicator was nucleopeptidic with information storage function carried out by the XNA and polymerase function carried out by the peptide.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Jeremy Tame, Andy Bates, and Arnout Voet for critical reading of the manuscript and Arnout Voet and Jan Zaucha for many constructive and critical discussions. This work was supported by RIKEN Initiative Research Funding to J.G.H. And funding from the Malopolska Centre of Biotechnology, awarded to J.G.H.

References

Angyan AF, Ortutay C, Gaspari Z. 2014. Are proposed early genetic codes capable of encoding viable proteins? *J Mol Evol*. 78(5):263–274.

Baaske P, Weinert FM, Duhr S, Lemke KH, Russell MJ, Braun D. 2007. Extreme accumulation of nucleotides in simulated hydrothermal pore systems. *Proc Natl Acad Sci USA*. 104(22):9346–9351.

Biro JC, Benyo B, Sansom C, Szlavecz A, Fordos G, Micsik T, Benyo Z. 2003. A common periodic table of codons and amino acids. *Biochem Biophys Res Commun*. 306(2):408–415.

Bromley EHC, Channon K, Moutevelis E, Woolfson DN. 2008. Peptide and protein building blocks for synthetic biology: from programming biomolecules to self-organized biomolecular systems. *ACS Chem Biol*. 3(1):38–50.

Cairns-Smith AG. 1982. Genetic takeover and the mineral origins of life. Cambridge University Press.

Carny O, Gazit E. 2005. A model for the role of short self-assembled peptides in the very early stages of the origin of life. *FASEB J*. 19(9):1051–1055.

Cech TR. 2012. The RNA worlds in context. *Cold Spring Harb Perspect Biol*. 4(7):a006742.

Da Silva L, Maurel MC, Deamer D. 2015. Salt-promoted synthesis of RNA-like molecules in simulated hydrothermal conditions. *J Mol Evol*. 80(2):86–97.

Dixit SB, Arora N, Jayaram B. 2000. How do hydrogen bonds contribute to protein–DNA recognition? *J Biomol Struct Dyn*. 17(Suppl 1):109–112.

Fishkis M. 2011. Emergence of self-reproduction in cooperative chemical evolution of prebiological molecules. *Origins Life Evol B*. 41(3):261–275.

Fletcher JM, Harniman RL, Barnes FR, Boyle AL, Collins A, Mantell J, Sharp TH, Antognozzi M, Booth PJ, Linden N. 2013. Self-assembling cages from coiled-coil peptide modules. *Science* 340(6132):595–599.

Fox SW, Harada K. 1958. Thermal copolymerization of amino acids to a product resembling protein. *Science* 128(3333):1214.

Giel-Pietraszuk M, Barciszewski J. 2006. Charging of tRNA with non-natural amino acids at high pressure. *FEBS J*. 273(13):3014–3023.

Giri V, Jain S. 2012. The origin of large molecules in primordial autocatalytic reaction networks. *PLoS ONE*. 7(1):e29546.

Herschy B, Whicher A, Camprubi E, Watson C, Dartnell L, Ward J, Evans JRG, Lane N. 2014. An origin-of-life reactor to simulate alkaline hydrothermal vents. *J Mol Evol*. 79(5–6):213–227.

Ikehara K. 2002. Origins of gene, genetic code, protein and life: comprehensive view of life systems from a GNC-SNS primitive genetic code hypothesis. *J Biosci*. 27(2):165–186.

Ikehara K. 2005. Possible steps to the emergence of life: the [GADV]-protein world hypothesis. *Chem Rec*. 5(2):107–118.

Illangasekare M, Sanchez G, Nickles T, Yarus M. 1995. Aminoacyl-RNA synthesis catalyzed by an RNA. *Science* 267(5198):643–647.

Issac R, Chmielewski J. 2002. Approaching exponential growth with a self-replicating peptide. *J Am Chem Soc*. 124(24):6808–6809.

Iyer L, Koonin E, Aravind L. 2003. Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. *BMC Struct Biol*. 3:1.

Keefe AD, Szostak JW. 2001. Functional proteins from a random-sequence library. *Nature* 410(6829):715–718.

Kemeny J. 1955. Man viewed as a machine. *Sci Am*. 192(4):58–68.

Knight RD, Landweber LF. 2000. Guilt by association: the arginine case revisited. *RNA* 6(4):499–510.

Kochavi E, Bar-Nun A, Fleminger G. 1997. Substrate-directed formation of small biocatalysts under prebiotic conditions. *J Mol Evol*. 45(4):342–351.

Koonin EV, Novozhilov AS. 2009. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life*. 61(2):99–111.

Koonin EV. 1991. The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. *J Gen Virol*. 72(9):2197–2206.

Kunin V. 2000. A system of two polymerases – a model for the origin of life. *Origins Life Evol B*. 30(5):459–466.

Kurland CG. 2010. The RNA dreamtime. *Bioessays* 32(10):866–871.

Lee DH, Granja JR, Martinez JA, Severin K, Ghadiri MR. 1996. A self-replicating peptide. *Nature* 382(6591):525–528.

Lehmann J, Reichel A, Buguin A, Libchaber A. 2007. Efficiency of a self-aminoacylating ribozyme: effect of the length and base-composition of its 3′ extension. *RNA* 13(8):1191–1197.
Leipe DD, Aravind L, Koonin EV. 1999. Did DNA replication evolve twice independently? *Nucleic Acids Res*. 27(17):3389–3401.
Leman L, Orgel L, Ghadiri MR. 2004. Carbonyl sulfide-mediated prebiotic formation of peptides. *Science* 306(5694):283–286.
Levy M, Ellington AD. 2003. Peptide-templated nucleic acid ligation. *J Mol Evol*. 56(5):607–615.
Liu Z, Beaufils D, Rossi JC, Pascal R. 2014. Evolutionary importance of the intramolecular pathways of hydrolysis of phosphate ester mixed anhydrides with amino acids and peptides. *Sci Rep*. 4:7440.
Manning MC, Illangasekare M, Woody RW. 1988. Circular dichroism studies of distorted alpha-helices, twisted beta-sheets, and beta turns. *Biophys Chem*. 31(1–2):77–86.
Martin W, Baross J, Kelley D, Russell MJ. 2008. Hydrothermal vents and the origin of life. *Nat Rev Microbiol*. 6(11):805–814.
McCleskey SC, Griffin MJ, Schneider SE, McDevitt JT, Anslyn EV. 2003. Differential receptors create patterns diagnostic for ATP and GTP. *J Am Chem Soc*. 125(5):1114–1115.
Miller SL. 1997. Peptide nucleic acids and prebiotic chemistry. *Nat Struct Biol*. 4(3):167–169.
Moore PB, Steitz TA. 2003. After the ribosome structures: how does peptidyl transferase work? *RNA* 9(2):155–159.
Morgens DW. 2013. The protein invasion: a broad review on the origin of the translational system. *J Mol Evol*. 77(4):185–196.
Nelson KE, Levy M, Miller SL. 2000. Peptide nucleic acids rather than RNA may have been the first genetic molecule. *Proc Natl Acad Sci USA*. 97(8):3868–3871.
Noller HF. 2012. Evolution of protein synthesis from an RNA world. *Cold Spring Harb Perspect Biol*. 4(4):a003681.
Patel BH, Percivalle C, Ritson DJ, Duffy CD, Sutherland JD. 2015. Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nat Chem*. 7(4):301–307.
Paul N, Joyce GF. 2004. Minimal self-replicating systems. *Curr Opin Chem Biol*. 8(6):634–639.
Pinheiro VB, Taylor AI, Cozens C, Abramov M, Renders M, Zhang S, Chaput JC, Wengel J, Peak-Chew SY, McLaughlin SH, et al. 2012. Synthetic genetic polymers capable of heredity and evolution. *Science* 336(6079):341–344.
Regan L, DeGrado WF. 1988. Characterization of a helical protein designed from first principles. *Science* 241(4868):976–978.
Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q, Baker D. 1997. Functional rapidly folding proteins from simplified amino acid sequences. *Nat Struct Biol*. 4(10):805–809.
Robertson MP, Joyce GF. 2012. The origins of the RNA world. *Cold Spring Harb Perspect Biol*. 4(5):a003608.
Rodin AS, Szathmary E, Rodin SN. 2011. On origin of genetic code and tRNA before translation. *Biol Direct*. 6:14.
Roviello G, Musumeci D, Castiglione M, Bucci EM, Pedone C, Benedetti E. 2009. Solid phase synthesis and RNA-binding studies of a serum-resistant nucleo-epsilon-peptide. *J Pept Sci*. 15(3):155–160.
Ruiz-Mirazo K, Briones C, de la Escosura A. 2014. Prebiotic systems chemistry: new perspectives for the origins of life. *Chem Rev*. 114(1):285–366.
Saladino R, Botta G, Pino S, Costanzo G, Di Mauro E. 2012. Genetics first or metabolism first? The formamide clue. *Chem Soc Rev*. 41(16):5526–5565.
Schimmel P, Henderson B. 1994. Possible role of aminoacyl-RNA complexes in noncoded peptide synthesis and origin of coded synthesis. *Proc Natl Acad Sci USA*. 91(24):11283–11286.
Schneider SE, O'Nei SN, Anslyn EV. 2000. Coupling rational design with libraries leads to the production of an ATP selective chemosensor. *J Am Chem Soc*. 122(3):542–543.
Sievers A, Beringer M, Rodnina MV, Wolfenden R. 2004. The ribosome as an entropy trap. *Proc Natl Acad Sci USA*. 101(21):7897–7901.
Tamura K. 2015. Beyond the frozen accident: glycine assignment in the genetic code. *J Mol Evol*. 81(3–4):69–71.
Tamura K, Schimmel P. 2003. Peptide synthesis with a template-like RNA guide and aminoacyl phosphate adaptors. *Proc Natl Acad Sci USA*. 100(15):8666–8669.
Trevino SG, Zhang N, Elenko MP, Luptak A, Szostak JW. 2011. Evolution of functional nucleic acids in the presence of nonheritable backbone heterogeneity. *Proc Natl Acad Sci USA*. 108(33):13492–13497.
Turk RM, Chumachenko NV, Yarus M. 2010. Multiple translational products from a five-nucleotide ribozyme. *Proc Natl Acad Sci USA*. 107(10):4585–4589.
van der Gulik P, Speijer D. 2015. How amino acids and peptides shaped the RNA world. *Life* 5(1):230.
Vetsigian K, Woese C, Goldenfeld N. 2006. Collective evolution and the genetic code. *Proc Natl Acad Sci USA*. 103(28):10696–10701.
Vidonne A, Philp D. 2009. Making molecules make themselves – the chemistry of artificial replicators. *Eur J Org Chem*. 2009(5):593–610.
Woese CR. 1965. On the evolution of the genetic code. *Proc Natl Acad Sci USA*. 54(6):1546–1552.
Wolf YI, Koonin EV. 2007. On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization. *Biol Direct*. 2:14.
Yarus M, Caporaso JG, Knight R. 2005. Origins of the genetic code: the escaped triplet theory. *Annu Rev Biochem*. 74:179–198.
Yarus M, Widmann JJ, Knight R. 2009. RNAamino acid binding: a stereochemical era for the genetic code. *J Mol Evol*. 69(5):406–429.
Yarus M. 2011. Getting past the RNA world: the initial darwinian ancestor. *Cold Spring Harbor Perspect Biol*. 3(4):a003590.
Zhang BL, Cech TR. 1997. Peptide bond formation by in vitro selected ribozymes. *Nature* 390(6655):96–100.