Title

Prebiotically plausible mechanisms increase compositional diversity of nucleic acid sequences

Journal

Nucleic Acids Research, 40(10)

ISSN

0305-1048

Authors

Derr, Julien
Manapat, Michael L
Rajamani, Sudha
et al.

Publication Date

2012-05-01

DOI

10.1093/nar/gks065

Published online 7 February 2012 *Nucleic Acids Research, 2012, Vol. 40, No. 10* ***4711–4722***
doi:10.1093/nar/gks065

Prebiotically plausible mechanisms increase compositional diversity of nucleic acid sequences

Julien Derr[1], Michael L. Manapat[2,3], Sudha Rajamani[1], Kevin Leu[1], Ramon Xulvi-Brunet[1], Isaac Joseph[1], Martin A. Nowak[3] and Irene A. Chen[1,*]

[1]FAS Center for Systems Biology, [2]School of Engineering and Applied Sciences and [3]Program for Evolutionary Dynamics, Harvard University, Cambridge, MA 02138, USA

Received October 14, 2011; Revised December 24, 2011; Accepted January 18, 2012

ABSTRACT

During the origin of life, the biological information of nucleic acid polymers must have increased to encode functional molecules (the RNA world). Ribozymes tend to be compositionally unbiased, as is the vast majority of possible sequence space. However, ribonucleotides vary greatly in synthetic yield, reactivity and degradation rate, and their non-enzymatic polymerization results in compositionally biased sequences. While natural selection could lead to complex sequences, molecules with some activity are required to begin this process. Was the emergence of compositionally diverse sequences a matter of chance, or could prebiotically plausible reactions counter chemical biases to increase the probability of finding a ribozyme? Our *in silico* simulations using a two-letter alphabet show that template-directed ligation and high concatenation rates counter compositional bias and shift the pool toward longer sequences, permitting greater exploration of sequence space and stable folding. We verified experimentally that unbiased DNA sequences are more efficient templates for ligation, thus increasing the compositional diversity of the pool. Our work suggests that prebiotically plausible chemical mechanisms of nucleic acid polymerization and ligation could predispose toward a diverse pool of longer, potentially structured molecules. Such mechanisms could have set the stage for the appearance of functional activity very early in the emergence of life.

INTRODUCTION

The biology of modern organisms is based on RNA, DNA and proteins, but this biochemistry was probably preceded by a stage in which RNA molecules acted both as chemical catalysts and carriers of genetic information. Evidence for this early stage of life (the 'RNA world') includes the similarity of ancient chemical cofactors to certain ribonucleotides and the discovery that the catalytic core of the ribosome is composed of RNA (1–5). Possible pathways for the prebiotically plausible synthesis of the components of RNA and the polymerization of ribonucleotides have been reported by several groups (3,6–8). While natural selection could enhance low catalytic activity, the very earliest ribozymes must have arisen through chemical processes (3). Understanding the details of this initial emergence is a deep conceptual puzzle (9).

The first ribozymes must have emerged from pools of short sequences that were low in diversity and information content, but it is unclear how the complexity of these pools could be increased (10,11). These sequence pools would have been limited for at least two reasons. First, monomers would have different abundances because they are synthesized and degraded by different pathways. For example, a concentrated eutectic phase solution of ammonium cyanide yields significantly more adenine than guanine, uracil and cytosine (roughly 10× or more) (6). Degradation affects the nucleobases differently, with cytosine being particularly susceptible to spontaneous deamination (12). Indeed, the abundances of nucleobases detected in meteorites also vary by one or more orders of magnitude (13–16). Second, the rate at which different monomers are polymerized can vary by an order of magnitude (6,17–20). In one study of montmorillonite-catalyzed RNA synthesis, this bias led to a large reduction in diversity, as only 3 out of 32 possible pentamer sequences were formed in detectable amounts from a mixture of activated A and C monomers (21). These biases reduce the diversity of the sequences generated, restricting exploration of sequence space and thus reducing the probability of generating a sequence with biological function. While compositional biases might increase the probability of generating functional RNA (22–24), the magnitude of the biases associated with prebiotically plausible polymerization is still substantially larger than potentially favorable biases. Interestingly, ribozymes with

*To whom correspondence should be addressed. Tel: +1 617 384 9647; Fax: +1 617 496 5425; Email: ichen@lsdiv.harvard.edu

limited compositional diversity have been made by a combination of rational design and *in vitro* evolution on restricted alphabets, but the resulting ribozymes exhibited decreased catalytic efficiency (the three-letter alphabet gave a 2500-fold decrease in k_{cat} relative to the four-letter alphabet; the two-letter alphabet gave a further 10-fold decrease in k_{cat} as well as a large decrease in total product conversion due to ribozyme misfolding) (25,26). While fine-tuning the conditions—e.g. by adjusting monomer ratios to counteract reduced reactivity or limiting UV irradiation to attain an appropriate monomer ratio (7,27,28)—could potentially overcome these biases, such conditions would be unlikely early on. Therefore, we sought more general mechanisms to counter compositional bias in nucleic acid pools undergoing prebiotically plausible reactions. Our experiments using DNA and simulations of binary sequences demonstrate that template-directed ligation is one such mechanism. Our RNA folding simulations suggest that compositionally diverse sequences are more likely to fold into stable structures compared with the substantially biased sequences that would be derived from template-independent processes. Greater stability is one of the factors promoting greater functional activity in RNA aptamers (29,30). In addition, our simulations indicate that template-directed ligation would shift the pool toward longer sequences, another important factor for activity (31,32). Our results suggest that a broad exploration of interesting sequence space was possible in prebiotic sequence pools despite initial chemical biases.

MATERIALS AND METHODS

Simulations

Two types of simulation were performed (stochastic and deterministic). Both simulations are limited for computational reasons. The stochastic simulation keeps track of each monomer or oligomer and implements a specific reaction at each time step. In practice, the simulation cannot keep track of an infinite number of reactants and possible reactions, so the system is limited by the total number of monomers considered. This mimics a protocell containing a relatively small number of monomers (e.g. 400, in which case the longest possible sequence would be 400 monomers). The stochastic simulation can generate sequences of ribozyme length. In contrast, the deterministic simulation keeps track of all possible species and reactions among them simultaneously, mimicking a very large, well-mixed system. In practice, the deterministic simulation cannot keep track of an infinite number of species, so the system must be truncated at a certain maximum length (e.g. 12).

The stochastic simulation was based on the Gillespie algorithm (33). Each simulation began with a pool of monomers. The total number of monomers in the system was typically limited to 400. During each iteration, an exponential waiting time was generated before a single reaction (concatenation, template-directed ligation or hydrolysis) occurred according to the relative rates of all possible reactions. Concatenation could occur between any two monomers or polymers. Realistic bias was introduced in the simulation as either a concatenation rate that depended on the identity of the 5′ monomer (e.g. reactivity ratio of 19:1), or as a difference in the initial number of monomers of each type (e.g. abundance ratio of 9:1). Template-directed ligation was possible if a 6-mer segment of one sequence (the template) was complementary to the trimer at the 3′-end of one substrate and the trimer at the 5′-end of the second substrate. Any sequence of length 6 or greater was a potential template; any sequence of length 3 or greater was a potential substrate. Circularization reactions were not considered. Hydrolysis of phosphodiester bonds occurred at a constant rate per bond. Characteristics of the system (average length and C_k) were measured at exponentially increasing time steps (i.e. 0, 1, 2, 4, 8, etc), and steady state was considered to be achieved when the characteristics at consecutive time steps deviated by <0.1%. SEs for these measurements were calculated from multiple simulation runs. The deterministic simulation used the same reactions as the stochastic simulation and kept track of the abundances of all possible sequences as the system evolved. The system was truncated at a maximum polymer length (typically 12) for computational tractability (i.e. polymers of the maximum length could not undergo further concatenation or ligation). Average C_k, average length and diversity were computed using steady-state abundances. The simulations were performed on the Odyssey Cluster of the FAS Research Computing Group at Harvard University. See Supplementary Data for simulation details.

Experiment: degenerate oligonucleotides for template-directed ligation of a heterogeneous sequence pool of DNA (four bases)

In heterogeneous pool reactions, all oligonucleotides were composed of A,C,G,T. Degenerate DNA oligonucleotides were obtained from Keck Oligo Synthesis Resource (Yale University, New Haven, CT, USA). Octamers and templates (40-mer) were synthesized as 5′-NNNN... using the facility's standard procedure for equimolar, degenerate oligonucleotides. Oligos were purified by reverse-phase cartridge. Octamers were phosphorylated as noted below, using non-radiolabeled ATP.

Experiment: template-directed chemical ligation reactions (four bases)

Reagents were purchased from Sigma-Aldrich (St Louis, MO, USA) unless otherwise specified. Synthetic degenerate oligonucleotides (templates of length 40; 5′-phosphorylated substrates of length 8) were mixed and ligated using cyanogen bromide following a previously published procedure (34). Reactions contained 1–2 μM DNA template and 16 μM DNA octamers. DNA sequences were mixed with buffer [0.23 M 2-(*N*-morpholino)ethanesulfonic acid, pH 7.4] and 19 mM $MgCl_2$, in 4.5 μl of aqueous solution, heated to 95° for 3 min and annealed by cooling on the benchtop for 15 min. The solution was placed on ice for 5 min and 0.5 μl CNBr (5 M in acetonitrile) was added (final

0.5 M). After 1 min, the reactions were quenched by 100 µl ethanol, ethanol precipitated for >1 h at −20° and centrifuged at 12 000*g* at 4° for 30 min. Supernatant was removed and the pellets were dried on the benchtop overnight. The pellets were resuspended in 50 µl of water and aliquots were mixed with loading buffer for electrophoresis through a 24% urea-polyacrylamide gel following a standard protocol.

Experiment: next-generation sequencing of templates and products of template-directed ligation (four bases)

Ligation products of length 16 were purified by urea–PAGE. Products were dephosphorylated with alkaline phosphatase following the manufacturer's protocol (all enzymes were obtained from New England Biolabs, Ipswich, MA, USA, unless otherwise specified), phenol–chloroform extracted and ethanol precipitated. Products, templates and octamers were resuspended and ligated to a barcoded 3′ adapter sequence for Illumina sequencing using T4 RNA ligase. The barcoded 3′ adapter sequences were (lowercase = RNA):

TAG1: 5′-ucgTGTCGTATGCCGTCTTCTGCTTGTddC,
TAG2: 5′-ucgCATCGTATGCCGTCTTCTGCTTGTddC,
3-TAG1: 5′-ucgTACTCGTATGCCGTCTTCTGCTTGTddC,
3-TAG2: 5′-ucgACATCGTATGCCGTCTTCTGCTTGTddC,
3-TAG3: 5′-ucgCTATCGTATGCCGTCTTCTGCTTGTddC.

Products were gel purified by urea–PAGE and phosphorylated by PNK following the manufacturer's protocol, phenol–chloroform extracted and ethanol precipitated. Phosphorylated products were then ligated to the 5′ adapter sequence for Illumina sequencing using T4 RNA ligase. The 5′ adapter sequence was: 5′-AATGA TACGGCGACCACCGACAGGTTCAGAGTTCTA Caguccgacgauc. Products were gel purified by urea–PAGE and reverse transcribed with SuperScript III RT (Invitrogen, Carlsbad, CA, USA). cDNA products were gel purified on urea–PAGE and directly sequenced on an Illumina (Solexa) Genome Analyzer II. Polymerase chain reaction was not performed in order to avoid bias due to differential PCR amplification. As controls, template and substrate DNA were also prepared for sequencing following the same procedure.

Experiment: sequence analysis (four bases)

Sequence reads of the appropriate length containing the appropriately barcoded 3′ adapter sequence were extracted. The compositional diversity of each sequence was calculated for subsequence length $k = 3$. To eliminate effects due to the total length of the sequences, length 16 was chosen for analysis. Products were already of this length. For templates (length 40), C_3 was calculated for all 16-mers contained in the template and the average was used as the C_3 of the template. For octamers (length 8), simulated 16-mer 'products' were generated *in silico* by randomly joining octamer sequences obtained experimentally, and the C_3 of these simulated random products was calculated.

Experiment: template-directed ligation with a single template (two bases)

Each reaction contained 1–2 µM DNA template and 16 µM random radiolabeled DNA octamers (Eurofins MWG Operon, Huntsville, AL, USA and Sigma-Aldrich, St Louis, MO, USA). In single-template reactions, the template oligonucleotides were composed of C and T while the octamers were composed of A and G. The templates used in the reactions of Figure 3b were:

5′-TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT,
5′-CTCTCTCTCTCTCTCTCTCTCTCTCTCTCT,
5′-TCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCT,
5′-CCTCTTCTCCTCTTCTCCTCTTCTCCTCTTCT,
5′-CTCTTCTTTTCTCCTCTTCTTTTCTCCTCCCC.

This set of templates utilizes a relatively small subset (25) of the theoretically possible octamers. These 25 octamers were mixed in equimolar ratio, phosphorylated and radiolabeled by T4 polynucleotide kinase (New England Biolabs, Ipswich, MA, USA) and added to the reaction, which was performed as described earlier. Band intensities were measured using a Typhoon TRIO Variable Mode Imager (Piscataway, NJ, USA). See Supplementary Data for more details.

DEFINITIONS AND MODEL

Measuring compositional bias for a given sequence

To measure the compositional bias, we use a definition inspired by grammar complexity and the Shannon entropy (35–37). For a sequence s of length L, we define $H_k(s)$ by

$$H_k(s) = -\sum_i p_i \log_2(p_i),$$

where i is the index of a unique string of length k ($k < L$; $i = 1$ to 2^k for a binary string) and p_i is the frequency of the ith k-mer within s. If $k = 1$, H_k is just the Shannon entropy. The H_k for various k describe the compositional bias of a sequence, and the H_k for $k > 1$ reveal information that H_1 by itself does not. Consider this hypothetical scenario: '0' tends to follow '1' and vice versa, such that most sequences are alternating (e.g. '01010101…'). In this case, H_1 would be high even though the sequence is not very diverse. However, H_2 is low, reflecting the lack of diversity on the '2-mer' scale, and so on for higher k.

The H_k have the advantage of being straightforward to compute and relevant for RNA folding, and measuring the diversity of potential template subsequences (see below). Like any simple measure, however, they suffer from some drawbacks. First, H_k for a fixed k can suggest high internal sequence diversity when the sequence is actually highly regular. Some well-known measures of complexity [e.g. the number of states in the smallest finite state machine accepting the sequence (38),

Table 1. Example RNA sequences, compositional diversity and folding energy predicted by Viennafold (41)

Sequence (50-mer)	C_4	E_m (kcal/mol)
GAUUGCCUAUAAACCAUUUAAUAGCCCAGCAACAUGCACAAAUGGGCGUU	1.0	–9.1
CAAAAAAAAAACAAUAAUAAAUGGACAAUAACCCCAUUAAAAGAAGAUAC	0.84	–3.0
AUAAAAAAAAAAGAAAAAUUACAUAAAAAAAAGAAAUAAAACAAAAUAAC	0.65	0
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	0	0
AUGGUAAGUUCCCAAGGCGGGUUGGAAGAGAUAUCAUAGGAGCUUGUCGU (GTP aptamer 10-59)	1.0	–11.9
AUAUAUAUAUAUAUAUAUAUAUAUAUAUAUAUAUAUAUAUAUAUAUAU	0.18	–20.6

In rows 2 and 3, adenine is 5- or 10-fold more abundant than the other bases, respectively, representing a realistic bias in composition (6). The fifth row contains the sequence of a known aptamer (42); other sequences were computer-generated. The sixth row illustrates an unusual sequence with low C_4 and low folding energy.

or algorithmic complexity, the length of the shortest program generating the sequence (39,40)] do not suffer from this limitation, but they are intractable to compute and do not incorporate information about the scale at which the underlying chemical processes are operating. Second, the calculation of H_k for $k > 1$ involves non-independent, overlapping sequences. This would be appropriate for measuring the diversity of potential ligation sites. However, one might imagine a scenario with much competition for binding to the same template (high concentrations of similar sequences); in that case, overlapping sites might interfere with one another and perhaps an alternative measure of diversity would be more appropriate. Nevertheless, we find H_k to have heuristic value for the analysis described below.

The maximum possible H_k ($H_k^{\max}$) grows with total sequence length L. To focus on composition rather than length, we define the relative compositional diversity C_k of a sequence to be the ratio $H_k(s)/H_k^{\max}(L)$. In addition, we only compare C_k for sequences of the same length. $C_k(s)$ measures the heterogeneity of strings of length k within sequence s. $C_k(s)$ is zero when all subsequences are identical, and $C_k(s)$ is one when all possible subsequences appear equally often in s (Table 1). For our purpose, if k is chosen to be too small, the compositional bias of longer k-mers is not captured, but if k is too large, C_k cannot distinguish well among different sequences. We generally used $k = 3$ except when analyzing long sequences (50-mer), when we used $k = 4$.

We find C_k to be a useful measure of the diversity of k-mers within sequence s, because values of C_k close to 1 would be desirable in the RNA world for at least two reasons. First, high C_k characterizes the vast majority of sequence space, because the number of different sequences corresponding to a particular composition is greater if the composition is more uniform (43). The total number of possible unique sequences varies approximately exponentially with C_k (Figure 1a). Biases in monomer composition and reactivity would decrease the average C_k and thus restrict the exploration of sequence space. For example, a 10-fold bias in composition decreases the average C_k from 0.94 to 0.43, which represents a severe restriction in sequence space given the exponential dependence (Supplementary Data). While sequence space might contain many potentially structured molecules (44), any search through sequence space for which the average C_k is low would under-represent or omit a large fraction of possible sequences. Therefore, high average C_k is desirable for finding rare, functional molecules.

Second, C_k appears to be correlated with RNA folding energy. Intuitively, an internally diverse sequence would present more independent opportunities for finding complementary regions compared to a repetitive sequence of the same length. A simple toy model for hairpin formation verifies this effect (Supplementary Data). Known ribozymes have nearly maximal C_4 (average C_4 of ribozyme sequence families of length 40–60 is 0.970 ± 0.013; Figure 1b, Supplementary Data) (45). On the other hand, some repetitive sequences have both low C_k and low folding energy [e.g. $(AU)_n$; Table 1]. It is conceivable that high C_k in ribozymes could be merely a reflection of the fact that many were isolated from pools of nearly random sequences, which have high average C_k. To understand the relationship between C_k and folding, we computed the minimum folding energy (E_m) for a large number of random RNA sequences (Figure 1c and Table 1). The fact that most random sequences have high C_k would cause a spurious correlation between C_k and E_m simply because a greater range of E_m is sampled by more sequences at high C_k. To determine the correlation between C_k and E_m independently of this effect, we binned the RNA sequences according to C_4 and analyzed an equal number of unique sequences in five bins centered between 0.6 and 1. We also limited our analysis to sequences with GC content of 40–60%, to avoid potential effects due to GC content alone. The folding energies are

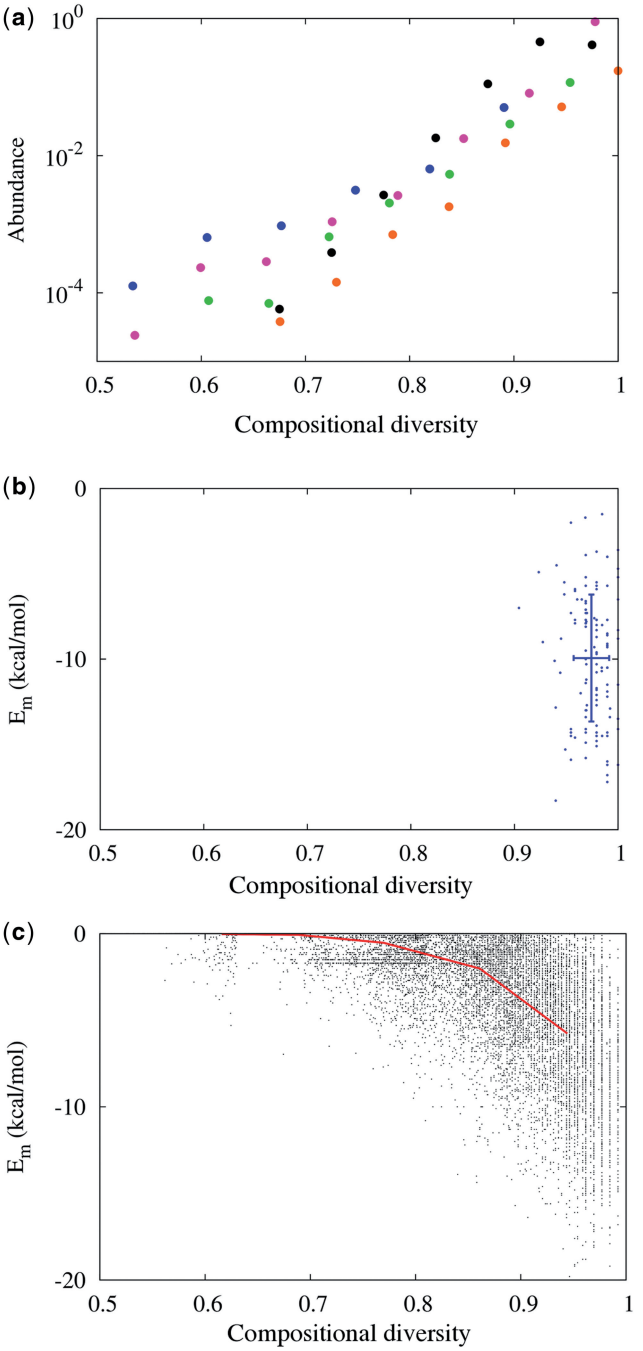

Figure 1. Compositional diversity, sequence space and predicted RNA folding energy. (**a**) Most of sequence space is of high compositional diversity. Histogram of C_4 for RNA sequences, computed from random sampling of 10^9 sequences of length 50 (black dots) *in silico*. The complete histogram for all possible sequences of shorter length is computable and is similar to that of the random sample of 50-mers (length 10 = blue, 12 = pink, 14 = green, 17 = orange). (**b**) Compositional diversity (C_4) and predicted minimum folding energy (E_m) for known ribozymes (length 40–60; see Supplementary Data) (45) are shown as blue dots with mean and SD (blue lines). (**c**) C_4 versus E_m (black dots) predicted by Viennafold (41) for 2.5×10^6 RNA sequences of length 50. To minimize effects from GC-content, we restricted the *in silico* sampling to sequences whose GC content is 40–60%. To avoid sampling artifacts, sequences were assigned to five bins according to C_4, and an equal number of unique sequences were analyzed in each bin. The bin averages are shown as the red line (see Supplementary Data for values and SDs).

correlated with C_4, and collectively the C_k explain 61% of the variance in E_m according to principal components regression analysis (Supplementary Data). There is a notable paucity of energetically stable, low C_4 sequences.

While stable folding is believed to be a prerequisite for function (46), C_k is not a perfect predictor of function. For example, sequence libraries containing deliberate repetitive patterns (alternating purine/pyrimidine) would have lower C_k on an average, but they perform at least as well as random libraries during SELEX because the design favors hairpin formation (46). Also, some rare structures might only be formed if the composition is biased. For example, sequences depleted in U and enriched in G are more likely to form stable structures, presumably because small regions of base-pairing are stabilized by this composition (22). Computational studies suggest that structures with long loop regions are favored by enrichment for A and C, and the optimal composition depends on the desired motif and structure (24). Loop regions in ribosomal RNAs tend to be A-rich, while G, C and U tend to comprise the stems (47). RNA folding simulations suggest that folding could be improved by a small compositional bias (nucleotide frequencies within 2-fold of each other) (23). It should be noted that the relationship of sequence, structure and function is complex and not yet fully understood, and well-folded structures are not always functional; mutations that preserve ribozyme fold might still destroy activity. Nevertheless, to the extent that structure may be important for function, in general C_k is a measure of compositional diversity and structural potential for a given sequence.

Measuring diversity of a pool

To quantify the diversity among different sequences in a pool, we also measure the population-level entropy D of a pool of molecules:

$$D = -\sum_{i=1}^{N} n_i \log_2(n_i),$$

where i is an index for unique sequences, N is the total number of unique sequences in the pool and n_i is the fraction of sequences in the pool that consist of copies of the ith sequence. D is zero if all molecules are identical and D is maximal when molecules are distributed uniformly through sequence space. In contrast to C_k (a property of each sequence), D is a property of the entire pool.

Simulations of prebiotically plausible reactions

To understand the effect of prebiotically plausible chemical reactions on compositional diversity, we first simulated a population of binary sequences undergoing three reactions: concatenation, template-directed ligation and hydrolysis. During concatenation, the 5′-end of one monomer (or polymer) reacts with the 3′-end of another monomer (or polymer) with rate constant k_{con}. This process first polymerizes monomers into oligomers, and later joins monomers and oligomers in

template-independent reactions. Template-directed ligation can occur when two oligomers anneal adjacent to one another on a template sequence, leading to the ligation of the two oligomers with rate constant k_{lig} (48). Template-directed ligation appears to be a general phenomenon, occurring with peptides, small molecules and nucleic acids (49–51), and it can greatly accelerate bond formation (52–54). Interestingly, template-directed ligation of oligonucleotides appears to be relatively unbiased compared to monomer polymerization, permitting the incorporation of nucleotides that are effectively unreactive as monomers (55). For the purpose of modeling, we assume that three or more adjacent 'Watson–Crick base-pairs' (i.e. 0's pairing with 1's in our two-letter model) are required for annealing (34,54,56,57). Finally, hydrolysis of phosphodiester bonds, an important process for RNA molecules, occurs in our model at a constant rate per bond (k_h).

Model parameters

The parameters of our simulations are the two dimensionless ratios: $r_{con} = k_{con}c_0/k_h$, where c_0 is the concentration of monomers, which gives the relative strength of concatenation and hydrolysis; and $r_{lig} = k_{lig}c_0/k_{con}$, which gives the relative strength of template-directed ligation and concatenation. In experiments, c_0 is usually in the millimolar range, r_{con} is roughly 1–100, and r_{lig} is between 10^3 and 10^7 (see Supplementary Data), so we use these parameters in our simulations. For computational tractability, we use a two-base system in the modeling for the purpose of building intuition. A two-base system has been proposed as a progenitor of the four-base system (12,58), and a ribozyme can be composed of only two bases (25). However, because a four-base system would be more realistic and it has been argued that this alphabet size is optimal (59), we also investigated the four-base system to the extent that it was computationally tractable. Regardless, a four-base system is used in our DNA-based experiments testing the predictions of the simulations.

Based on these chemical reactions, we implemented a stochastic simulation of a small reactor (e.g. a protocell) and a deterministic simulation mimicking a very large reactor. The stochastic simulations were initiated with ~400 monomers (corresponding to a concentration of ~10 mM in a protocell ~50–100 nm in diameter). We recorded the average C_k, D and the length distribution after the reactors reached steady state (Supplementary Data). While the stochastic results are most relevant to prebiotic protocells, we used the deterministic results to understand diversity for computational reasons. The deterministic simulations are generalizations of a previously described 'prelife' framework (Supplementary Data) (60–62). We examined both possible sources of bias: (i) biased reactivity and (ii) biased initial monomer abundance. Based on the reactivity and abundance differences of the literature cited earlier, the bias examined in each case was roughly one order of magnitude.

RESULTS

Simulation: template-directed ligation causes a shift toward longer sequences

In the absence of template-directed ligation, both sources of bias resulted in an exponential relationship between length and abundance at steady state, with the scaling determined by the ratio r_{con}. We give an analytical proof of this relationship, which has been seen in other models of polymerization (63), in Supplementary Data. While increasing concatenation would also create longer products due to greater bond formation, even high concatenation rates would still give an exponentially decreasing distribution of lengths. In contrast, template-directed ligation skewed the distribution qualitatively toward longer lengths (Figure 2a and Supplementary Data), resulting in a substantial excess of long sequences compared to an exponential distribution. The skew may occur because this process uses somewhat long substrates (greater than or equal to three bases) to make longer products in relatively few steps. Since ribozymes and

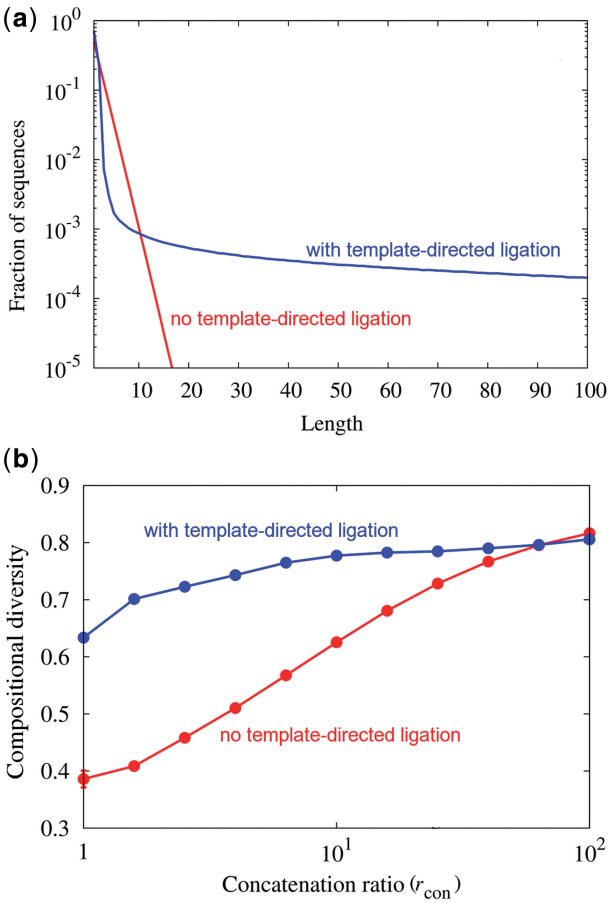

Figure 2. Template-directed ligation increases average length and compositional diversity *in silico*. (**a**) Length distribution of binary sequences with or without template-directed ligation [r_{lig} = 0 (red) or 10^6 (blue); r_{con} = 10 in both cases]. Length is the number of bases per molecule. (**b**) Compositional diversity C_3 at several r_{con} values, with or without template-directed ligation, when monomer reactivity is biased [19-fold difference between k_{con}; r_{lig} = 0 (red) or 10^6 (blue); length = 15].

aptamers typically have a length of 30 bases or greater (45), template-directed ligation could improve the chance of obtaining functional molecules simply by increasing the number of long polymers. For example, as r_{lig} increased from 0 to 10^6 ($r_{con} = 10$), the mass fraction of ribozyme-length sequences (>30 bases) increased from 0% (numerically undetectable) to >5%. A similar trend is seen using a four-base simulation (Supplementary Data).

Simulation: template-directed ligation increases compositional diversity

Template-directed ligation increased the average compositional diversity C_3 beyond that achieved by concatenation alone when analyzing product sequences of the same length. When reactivities were biased, C_3 depended on the rates of both concatenation and template-directed ligation. Without template-directed ligation, a system with higher r_{con} had higher average C_3 (Figure 2b). This effect appears to be a consequence of mass action, as the less reactive monomer is increasingly incorporated into polymers when concatenation is fast relative to hydrolysis. A simplified analytical model demonstrates this effect (Supplementary Data). However, at a given r_{con}, template-directed ligation further increased average C_3 (Figure 2b), particularly at low concatenation rates. A four-base system appears to give similar results, with the caveat that our analysis was limited by computational tractability (Supplementary Data). Since template-directed ligation, like concatenation, increased the amount of bond formation relative to hydrolysis, one possible explanation might again be mass action. Therefore, we also measured C_3 as a function of the total rate of bond-forming events (concatenation and template-directed ligation). At the same rate of bond formation, template-directed ligation still increased the C_3 of the sequences, such that the majority of the increase of C_3 with r_{lig} was not simply the result of increased bond formation (Supplementary Data). Another possible mechanism for this increase is that template-directed ligation is a relatively unbiased mode of ligation compared to concatenation. To illustrate this point, we performed simulations in which we artificially relaxed the requirement for complementary base-pairing in template-directed ligation, which we call 'relaxed-ligation'. Relaxed-ligation retains the effect of introducing unbiased reactions while eliminating the more subtle effects stemming from the information content of the reacting sequences. Relaxed-ligation produces average C_3 that are similar to comparable template-directed ligation simulations (Supplementary Data). This result suggests that the unbiased nature of template-directed ligation is an important explanation for the increase of C_3 with template-directed ligation.

In simulations where monomer abundance was biased, concatenation alone resulted in relatively low average C_3 (~0.4), independent of r_{con}. The effect of template-directed ligation was complicated but it tended to increase compositional diversity (Supplementary Data). This may be due to increased incorporation of the less abundant monomer through complementary base pairing in addition to the effects detailed below. Overall, both sets of simulations suggested that template-directed ligation caused a relative increase of compositionally diverse sequences.

Simulation: diversity of the pool

While C_k measures internal heterogeneity in a sequence and D measures population-wide diversity, we found that these measures were highly correlated in our simulations across a range of parameters (Supplementary Data). Therefore, increased average compositional diversity within a sequence implied increased diversity among molecules in the pool.

Experimental: compositional diversity in template-directed ligation of DNA

To experimentally test our main prediction that template-directed ligation increases the average compositional diversity of a heterogeneous pool of sequences containing all 4 nt, we performed template-directed ligation in a pool of sequences made by degenerate DNA oligonucleotide synthesis with all four bases (A, C, G, T). The experiments described here were performed with DNA (not RNA). We chose to use DNA because reactivity differences during synthesis are well known (64). Slight differences in phosphoramidite reactivity result in small biases in the composition of a degenerate pool, analogous to the larger biases resulting from reactivity differences in prebiotic syntheses. Since the reactivity biases in phosphoramidite synthesis are relatively small, this experiment is a stringent test of whether template-directed ligation can increase average C_k. We studied whether the bias would be countered by template-directed ligation. Degenerate templates (length 40; four bases) and octamer substrates (four bases) were used to perform ligation. The size difference between the templates, substrates and expected ligation products permitted later gel purification of the products. Bond formation was catalyzed by cyanogen bromide after an annealing step (34) and the products of ligation were isolated by gel purification. Templates, octamers and products were sequenced using the Illumina platform (Supplementary Data). We measured the proportion of sequences that had C_3 close to that of ribozymes (C_3 of 0.95 or greater) and found that the ligation products were significantly shifted toward higher C_3 compared to the templates (Figure 3a and Supplementary Data). Ligation products also had higher C_3 than sequences predicted from random concatenation of the sequenced octamers, indicating that template-directed ligation could increase average C_3 beyond unbiased concatenation by favoring compositionally diverse templates. The proportion of ligation products of high C_3 was similar to that of a uniform random pool, indicating that template-directed ligation quantitatively countered the initial bias of synthesis (Figure 3a). The shift was not due to artifacts from comparing samples of different length, experimental bias during sequencing or a shift in GC content (Supplementary Data). We attempted to ascertain whether sequence elements from a recently

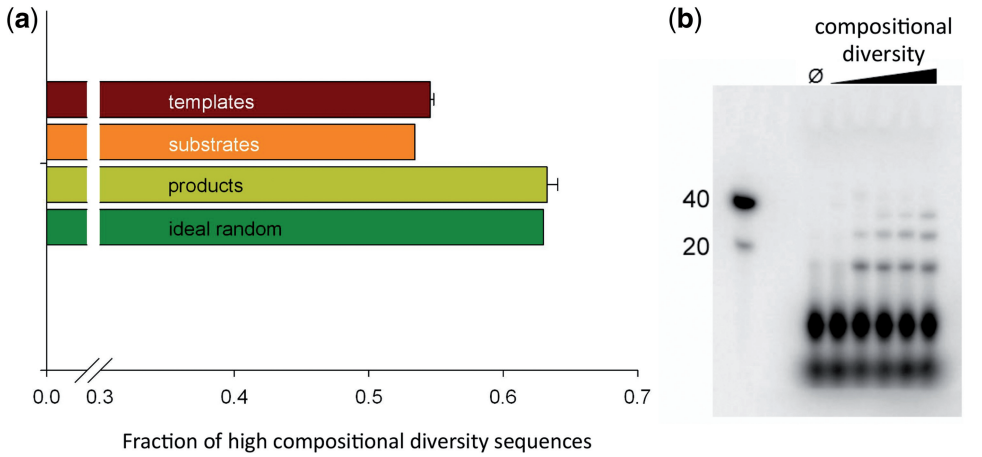

Figure 3. Experimental relationship between compositional diversity and template-directed ligation. (**a**) Fraction of DNA sequences having high compositional diversity ($C_3 > 0.95$; analysis length = 16) after template-directed ligation in a heterogeneous pool of degenerate oligonucleotides including four bases. A greater fraction of reaction products (yellow) have high C_3 relative to the templates (red = average C_3 of 16-mers contained in sequenced 40-mer templates) and substrates (orange = C_3 of 16-mers from *in silico* non-templated, random concatenation of experimentally sequenced octamers). The fraction of high C_3 sequences in a uniform random pool is shown in green. Error bars are SDs from replicate sequencing experiments. (**b**) Polyacrylamide gel showing higher molecular weight products of template-directed ligation for different single templates from a binary alphabet. Molecular weight markers are given in the left lane. '**Ø**' indicates a reaction without template added. Template C_3 increases from left ($C_3 = 0$) to right ($C_3 = 0.97$; see 'Materials and Methods' section for list of sequences).

described RNA replicase (65) could be found at greater frequency in the pool of template-directed ligation products compared with random concatenation of the octamers; no difference was apparent by this test (Supplementary Data), but the importance of this finding is tempered by our incomplete understanding of the sequence elements supporting ribozyme function.

The error of our measurements of compositional diversity in ligation reactants and products could be calculated in two ways: (i) SD among experiments, as shown in Figure 3a or (ii) SD from bootstrapping subsamples. The error (i) reflects the deviation between experiments. The error (ii) reflects the sampling error of sequencing. The sampling error is similar in magnitude to the error between experiments (Supplementary Data). Also, a possible source of differences between templates, octamers and ligation products is bias introduced by RNA ligase during preparation for deep sequencing. This bias is most pronounced at the 5′- and 3′-ends of the sequence reads (66). To confirm that the differences in measured C_k among these samples were not due to artifacts from bias at the ends of the sequences, we randomized the first and last base of each sequence read (i.e. replaced the 3′ and 5′ bases with a randomly chosen base: A,C,G,T). The C_k of this end-randomized set of sequences were calculated. We found that end-randomization did not affect the conclusion that ligation products had significantly higher C_k than the templates and octamers (Supplementary Data).

These results confirmed the prediction that template-directed ligation increases the compositional diversity of a heterogeneous pool of DNA sequences comprising four bases. One possible mechanism for this increase could be that internally diverse sequences were better templates. We can calculate the ratio R of the probabilities of template-directed ligation occurring on a high C_k (p_{high}) versus low C_k (p_{low}) template:

$$R = \frac{p_{high}}{p_{low}} = \frac{1-(1-p_1p_2)^N}{p_1 \times (1-(1-p_2)^N)}$$

where p_1 is the probability of the template annealing to two adjacent fragments and p_2 is the probability of bond formation. Both p_1 and p_2 are <1 in practical situations. The ratio R is therefore always >1 (Supplementary Data), meaning that high C_k sequences are more likely to be templates. Therefore, templates with high C_k should be more likely to propagate their sequence information.

To test this experimentally, we studied the dependence of ligation efficiency on C_3 for different templates. A set of binary DNA templates (two bases: C,T; 32-mer) of varying C_3 was designed such that any 8-mer subsequence within the templates was 1 of 25 known octamer sequences (two bases: A,G), allowing the use of a defined set of substrates. Although this base composition is not a good mimic of a prebiotic reaction, it was chosen to minimize intramolecular secondary structure in order to focus on the effect of sequence heterogeneity. The templates were mixed with an excess of radiolabeled 5′-phosphorylated binary DNA octamers (A,G; Supplementary Data). All sequences had a GC content of 47–50% except for the template with $C_3 = 0$. Ligation reactions were analyzed by polyacrylamide gel electrophoresis to visualize higher molecular weight products. We found a positive relationship between C_3 and amount of products formed (Figure 3b). A similar trend was observed using random degenerate octamers and longer templates (Supplementary Data). This suggests that ligation efficiency on heterogeneous templates is one mechanism by which C_k increases in the products of template-directed ligation.

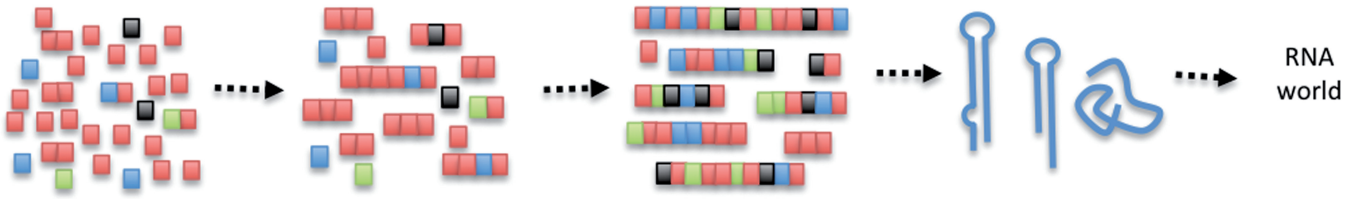

Figure 4. Proposed prebiotic scenario. Monomers first concatenate into compositionally biased short oligomers. When the oligomers are long enough to act as templates, template-directed ligation produces relatively long, compositionally diverse sequences. These sequences can fold into stable structures, some of which may be catalytically active, leading to the RNA world.

DISCUSSION

Our simulations and experiments suggest that compositional diversity in a pool of nucleic acids could have emerged early on, despite biases in monomer abundance and reactivity. The increase due to template-directed ligation may have at least two causes. First, template-directed ligation is relatively unbiased compared to concatenation, so it would counter the intrinsic bias of the system. Second, ligation may happen more efficiently on a compositionally diverse template, because internal sequence correlations in a low C_k template reduce the number of independent possibilities for ligation. Essentially, a compositionally diverse template could utilize a greater fraction of the substrate pool while the subsequences within a repetitive template would compete with each other for substrates.

While it may be possible to imagine scenarios where different biases cancel one another to produce complex pools, such finely tuned rates are unlikely in real systems. Based on our results using two-letter simulations and template-dependent DNA ligation, we can suggest the following prebiotic scenario (Figure 4). Initially the sequence pool would consist of short, highly compositionally biased sequences resulting from differences in reactivity among the nucleotides. However, once these sequences become long enough to serve as templates (length ≥ 6), the general mechanism of template-directed ligation would favor propagation of internally diverse sequences. A small increase in diversity would correspond to an exponentially large increase in the fraction of sequence space that would be explored. One should note that the optimal compositional diversity for forming secondary structures may be less than the maximum possible. For example, compositional bias toward GC-rich sequences may be a reasonable criterion for identifying non-coding RNAs in the genome due to the effect on folding stability (67), although formal structure itself does not appear to be a good criterion (perhaps because the compositional diversity of a genome tends to be fairly high, giving a relatively large probability of forming structures). Nevertheless, it is unclear whether the bias from prebiotic processes would be in the correct direction to favor structures, and it would still be desirable to increase diversity over the substantial ~10-fold initial compositional bias estimated for prebiotic reactions. As the compositional bias disappeared, well-folded sequences could emerge. In addition, template-directed ligation would rapidly stitch together short sequences to produce a qualitative shift toward long sequences. The combined effects on diversity and length would enable the generation of ribozymes. These first inefficient ribozymes could then 'jump-start' the evolution of sequences with greater function and complexity through natural selection, especially within a spatially restricted context (44,63,68).

Although it has been previously hypothesized that abstract measures of primary sequence information are unrelated to RNA function (11,69), we found a correlation between folding energy and C_k *in silico*. This suggests that internal heterogeneity, which is calculable from primary sequence alone, is an interesting measurement in addition to functional information or genomic complexity (which require knowledge of functional activity or fitness, respectively) (36,69). An important caveat regarding this relationship is that the minimum free energy of a sequence is only one of several features that would be important for functional activity. Other features would also be desirable [e.g. a large energy gap between the most stable fold and misfolded structures, or low structural 'plasticity' (70)]. Many desirable features are poorly understood. Interestingly, our simulations results suggest that, although template-independent processes result in exponential length distributions, template-directed ligation would skew the distribution qualitatively toward long sequences. A minimum length appears to be required to find certain activities. This was demonstrated by a series of selections for isoleucine aptamers that differed only by the length of the random region; no aptamers were isolated at the shortest length (16 bases) (32). Therefore, template-directed ligation may increase the probability of finding functional molecules by increasing the frequency of long sequences. Our results also highlight the importance of templating as a special property of nucleic acids during the origin of life: in addition to enabling the faithful replication of information, the ability to template could have promoted a search of functionally rich regions of sequence space.

Our modeling, while adequate for generating a hypothesis that could be experimentally tested, could be made more realistic in several ways. The alphabet size could be increased to include four bases, although this modification is computationally expensive because longer sequences would be needed to differentiate among the more varied compositions. Secondary structure could be included, which might interfere with templating and protect against degradation (71). The fact that our DNA experiments (including a four-letter alphabet and the

possibility of secondary structure) show that template-directed ligation increases compositional diversity suggests that the overall result is not greatly influenced by the simplifications in the modeling. The rate of non-templated concatenation might depend on the length of the reactants, and circularization reactions might select against certain lengths. Interestingly, models of prebiotic polymerization of monomers that neglect the concatenation of oligomers (i.e. strong dependence of rate on substrate length) also yield exponential length distributions at equilibrium (61–63,72), suggesting that the qualitative distribution is not changed by the inclusion of oligomer concatenation. More experimental investigation would be required to understand the possible dependencies. Advances in the modeling based on such realistic features would be worthwhile for a more detailed prediction of the outcome of prebiotically plausible reactions. Our experiments used DNA because biases during synthesis are well known and we expect DNA to resemble RNA with respect to annealing of oligonucleotides. While the details of formation of secondary structure differ for DNA and RNA, both types of molecule can fold into catalytically active structures (73). Nevertheless, an investigation of compositional diversity using RNA would be more realistic as a model of the RNA world.

It is generally assumed that the emergence of ribozymes was the result of natural selection for replication in a pool of RNA sequences (3,44), but the pathway for generating the very first ribozymes is unknown. The probability of finding functional sequences also depends on the desired activity. Functions that can be performed by short motifs, such as aminoacylation or self-cleavage (24,31,74), could potentially be found even in highly biased sequence pools since the probability of finding the motif is relatively large. However, more sophisticated functions that appear to require longer motifs, such as an RNA polymerase (27,65), would be more likely to arise in a sequence pool if the prebiotic compositional bias were at least somewhat mitigated. It has been suggested that physical ordering effects alone could not produce functional molecules (11). Here, we have shown how long, compositionally diverse and well-folded sequences might be produced as a consequence of prebiotically plausible chemical mechanisms.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–24 and Supplementary References [75–86].

ACKNOWLEDGEMENTS

We thank Niles Lehman, Jeffrey Bada, Jim Collins, Allan Drummond, Suckjoon Jun, Arthur Lander, David Liu, Andrew Murray, Eugene Shakhnovich and Jack Szostak for comments.

FUNDING

Human Frontiers Science Program (postdoctoral fellowships to J.D. and R.X.); Harvard University (I.A.C., Bauer Fellow); NIH grant GM068763 to the Center for Modular Biology at Harvard; NSF/NIH Joint Program in Mathematical Biology (NIH grant R01GM078986 to M.A.N.); John Templeton Foundation (M.A.N.); Bill and Melinda Gates Foundation (Grand Challenges grant 37874 to M.A.N.) and J. Epstein (M.A.N.). Funding for open access charge: NIH (grant GM068763).

Conflict of interest statement. None declared.

REFERENCES

1. Crick,F.H. (1968) The origin of the genetic code. *J. Mol. Biol.*, **38**, 367–379.
2. Orgel,L.E. (1968) Evolution of the genetic apparatus. *J. Mol. Biol.*, **38**, 381–393.
3. Orgel,L.E. (2004) Prebiotic chemistry and the origin of the RNA world. *Crit. Rev. Biochem. Mol. Biol.*, **39**, 99–123.
4. Woese,C.R., Dugre,D.H., Dugre,S.A., Kondo,M. and Saxinger,W.C. (1966) On the fundamental nature and evolution of the genetic code. *Cold Spring Harb. Symp. Quant. Biol.*, **31**, 723–736.
5. Nissen,P., Hansen,J., Ban,N., Moore,P.B. and Steitz,T.A. (2000) The structural basis of ribosome activity in peptide bond synthesis. *Science*, **289**, 920–930.
6. Miyakawa,S., Cleaves,H.J. and Miller,S.L. (2002) The cold origin of life: B. Implications based on pyrimidines and purines produced from frozen ammonium cyanide solutions. *Orig. Life Evol. Biosph.*, **32**, 209–218.
7. Powner,M.W., Gerland,B. and Sutherland,J.D. (2009) Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature*, **459**, 239–242.
8. Rajamani,S., Vlassov,A., Benner,S., Coombs,A., Olasagasti,F. and Deamer,D. (2008) Lipid-assisted synthesis of RNA-like polymers from mononucleotides. *Orig. Life Evol. Biosph.*, **38**, 57–74.
9. Davies,P. (2001) The origin of life. II: how did it begin? *Sci. Prog.*, **84**, 17–29.
10. Joyce,G.F. (1987) Nonenzymatic template-directed synthesis of informational macromolecules. *Cold Spring Harb. Symp. Quant. Biol.*, **52**, 41–51.
11. Abel,D.L. and Trevors,J.T. (2005) Three subsets of sequence complexity and their relevance to biopolymeric information. *Theor. Biol. Med. Model.*, **2**, 29.
12. Levy,M. and Miller,S.L. (1998) The stability of the RNA bases: implications for the origin of life. *Proc. Natl Acad. Sci. USA*, **95**, 7933–7938.
13. Glavin,D. and Bada,J. (2004) *35th Lunar and Planetary Science Conference*. League City, TX, pp. 1022.
14. Shimoyama,A., Hagishita,S. and Harada,K. (1990) Search for nucleic-acid bases in carbonaceous chondrites from Antarctica. *Geochem. J.*, **24**, 343–348.
15. Stoks,P.G. and Schwartz,A.W. (1979) Uracil in carbonaceous meteorites. *Nature*, **282**, 709–710.
16. Stoks,P.G. and Schwartz,A.W. (1981) Nitrogen-heterocyclic compounds in meteorites - significance and mechanisms of formation. *Geochim. Cosmochim. Acta*, **45**, 563–569.
17. Kawamura,K. and Ferris,J.P. (1999) Clay catalysis of oligonucleotide formation: kinetics of the reaction of the 5′-phosphorimidazolides of nucleotides with the non-basic heterocycles uracil and hypoxanthine. *Orig. Life Evol. Biosph.*, **29**, 563–591.
18. Sawai,H. and Orgel,L.E. (1975) Letter: oligonucleotide synthesis catalyzed by the Zn-2+ ion. *J. Am. Chem. Soc.*, **97**, 3532–3533.
19. Rajamani,S., Ichida,J.K., Antal,T., Treco,D.A., Leu,K., Nowak,M.A., Szostak,J.W. and Chen,I.A. (2010) Effect of stalling after mismatches on the error catastrophe in nonenzymatic nucleic acid replication. *J. Am. Chem. Soc.*, **132**, 5880–5885.

20. Ertem,G., Hazen,R.M. and Dworkin,J.P. (2007) Sequence analysis of trimer isomers formed by montmorillonite catalysis in the reaction of binary monomer mixtures. *Astrobiology*, **7**, 715–722.
21. Miyakawa,S. and Ferris,J.P. (2003) Sequence- and regioselectivity in the montmorillonite-catalyzed synthesis of RNA. *J. Am. Chem. Soc.*, **125**, 8202–8208.
22. Stich,M., Briones,C. and Manrubia,S.C. (2008) On the structural repertoire of pools of short, random RNA sequences. *J. Theor. Biol.*, **252**, 750–763.
23. Kennedy,R., Lladser,M.E., Wu,Z., Zhang,C., Yarus,M., De Sterck,H. and Knight,R. (2010) Natural and artificial RNAs occupy the same restricted region of sequence space. *RNA*, **16**, 280–289.
24. Knight,R., De Sterck,H., Markel,R., Smit,S., Oshmyansky,A. and Yarus,M. (2005) Abundance of correctly folded RNA motifs in sequence space, calculated on computational grids. *Nucleic Acids Res.*, **33**, 5924–5935.
25. Reader,J.S. and Joyce,G.F. (2002) A ribozyme composed of only two different nucleotides. *Nature*, **420**, 841–844.
26. Rogers,J. and Joyce,G.F. (1999) A ribozyme that lacks cytidine. *Nature*, **402**, 323–325.
27. Johnston,W.K., Unrau,P.J., Lawrence,M.S., Glasner,M.E. and Bartel,D.P. (2001) RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science*, **292**, 1319–1325.
28. Muller,U.F. (2006) Re-creating an RNA world. *Cell. Mol. Life Sci.*, **63**, 1278–1293.
29. Carothers,J.M., Oestreich,S.C. and Szostak,J.W. (2006) Aptamers selected for higher-affinity binding are not more specific for the target ligand. *J. Am. Chem. Soc.*, **128**, 7929–7937.
30. Carothers,J.M., Davis,J.H., Chou,J.J. and Szostak,J.W. (2006) Solution structure of an informationally complex high-affinity RNA aptamer to GTP. *RNA*, **12**, 567–579.
31. Knight,R. and Yarus,M. (2003) Finding specific RNA motifs: function in a zeptomole world? *RNA*, **9**, 218–230.
32. Legiewicz,M., Lozupone,C., Knight,R. and Yarus,M. (2005) Size, constant sequences, and optimal selection. *RNA*, **11**, 1701–1709.
33. Gillespie,D. (1977) Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, **81**, 2340–2361.
34. James,K.D. and Ellington,A.D. (1997) Surprising fidelity of template-directed chemical ligation of oligonucleotides. *Chem. Biol.*, **4**, 595–605.
35. Shannon,C. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.
36. Adami,C., Ofria,C. and Collier,T.C. (2000) Evolution of biological complexity. *Proc. Natl Acad. Sci. USA*, **97**, 4463–4468.
37. Lehman,N., Donne,M.D., West,M. and Dewey,T.G. (2000) The genotypic landscape during in vitro evolution of a catalytic RNA: implications for phenotypic buffering. *J. Mol. Evol.*, **50**, 481–490.
38. Sipser,M. (2005) *Introduction to the Theory of Computation*, 2nd edn. Thomson Course Technology, Boston.
39. Chaitin,G.J. (1975) Theory of program size formally identical to information-theory. *J. ACM*, **22**, 329–340.
40. Chaitin,G.J. (1977) Algorithmic information-theory. *IBM J. Res. Dev.*, **21**, 350–359.
41. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
42. Carothers,J.M., Oestreich,S.C., Davis,J.H. and Szostak,J.W. (2004) Informational complexity and functional activity of RNA structures. *J. Am. Chem. Soc.*, **126**, 5130–5137.
43. Stich,M. and Manrubia,S.C. (2011) Motif frequency and evolutionary search times in RNA populations. *J. Theor. Biol.*, **280**, 117–126.
44. Briones,C., Stich,M. and Manrubia,S.C. (2009) The dawn of the RNA World: toward functional complexity through ligation of random RNA oligomers. *RNA*, **15**, 743–749.
45. Lee,J.F., Hesselberth,J.R., Meyers,L.A. and Ellington,A.D. (2004) Aptamer database. *Nucleic Acids Research*, **32**, D95–100.
46. Ruff,K.M., Snyder,T.M. and Liu,D.R. (2010) Enhanced functional potential of nucleic acid aptamer libraries patterned to increase secondary structure. *J. Am. Chem. Soc.*, **132**, 9453–9464.
47. Gutell,R.R., Cannone,J.J., Shang,Z., Du,Y. and Serra,M.J. (2000) A story: unpaired adenosine bases in ribosomal RNAs. *J. Mol. Biol.*, **304**, 335–354.
48. Li,X. and Liu,D.R. (2004) DNA-templated organic synthesis: nature's strategy for controlling chemical reactivity applied to synthetic molecules. *Angew. Chem. Int. Ed. Engl.*, **43**, 4848–4870.
49. Lee,D.H., Granja,J.R., Martinez,J.A., Severin,K. and Ghadiri,M.R. (1996) A self-replicating peptide. *Nature*, **382**, 525–528.
50. Naylor,R. and Gilham,P.T. (1966) Studies on some interactions and reactions of oligonucleotides in aqueous solution. *Biochemistry*, **5**, 2722–2728.
51. Tjivikua,T., Ballester,P. and Rebek,J. (1990) Self-replicating system. *J. Am. Chem. Soc.*, **112**, 1249–1250.
52. Kanavarioti,A. and White,D.H. (1987) Kinetic analysis of the template effect in ribooligoguanylate elongation. *Orig. Life Evol. Biosph.*, **17**, 333–349.
53. Rohatgi,R., Bartel,D.P. and Szostak,J.W. (1996) Nonenzymatic, template-directed ligation of oligoribonucleotides is highly regioselective for the formation of 3′-5′ phosphodiester bonds. *J. Am. Chem. Soc.*, **118**, 3340–3344.
54. Sievers,D. and von Kiedrowski,G. (1994) Self-replication of complementary nucleotide-based oligomers. *Nature*, **369**, 221–224.
55. Ninio,J. and Orgel,L.E. (1978) Heteropolynucleotides as templates for non-enzymatic polymerizations. *J. Mol. Evol.*, **12**, 91–99.
56. Sawai,H., Totuka,S. and Yamamoto,K. (1997) Helical structure formation between complementary oligonucleotides. Minimum chain length required for the template-directed synthesis of oligonucleotides. *Orig. Life Evol. Biosph.*, **27**, 525–533.
57. Sawai,H. and Wada,M. (2000) Nonenzymatic template-directed condensation of short-chained oligouridylates on a poly(A) template. *Orig. Life Evol. Biosph.*, **30**, 503–511.
58. Wachtershauser,G. (1988) An all-purine precursor of nucleic acids. *Proc. Natl Acad. Sci. USA*, **85**, 1134–1135.
59. Szathmary,E. (1992) What is the optimum size for the genetic alphabet? *Proc. Natl Acad. Sci. USA*, **89**, 2614–2618.
60. Manapat,M., Ohtsuki,H., Burger,R. and Nowak,M.A. (2009) Originator dynamics. *J Theor. Biol.*, **256**, 586–595.
61. Manapat,M.L., Chen,I.A. and Nowak,M.A. (2010) The basic reproductive ratio of life. *J. Theor. Biol.*, **263**, 317–327.
62. Nowak,M.A. and Ohtsuki,H. (2008) Prevolutionary dynamics and the origin of evolution. *Proc. Natl Acad. Sci. USA*, **105**, 14924–14927.
63. Wu,M. and Higgs,P.G. (2009) Origin of self-replicating biopolymers: autocatalytic feedback can jump-start the RNA world. *J. Mol. Evol.*, **69**, 541–554.
64. Pollard,J., Bell,S.D. and Ellington,A.D. (2000) Design, synthesis, and amplification of DNA pools for construction of combinatorial pools and libraries. *Curr. Protoc. Mol. Biol.*, Chapter 24, Unit 24.2.
65. Wochner,A., Attwater,J., Coulson,A. and Holliger,P. (2011) Ribozyme-catalyzed transcription of an active ribozyme. *Science*, **332**, 209–212.
66. Eun,H.-M. (1996) *Enzymology Primer for Recombinant DNA Technology*. Academic Press, San Diego.
67. Rivas,E. and Eddy,S.R. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.
68. Szabó,P., Scheuring,I., Czárán,T. and Szathmary,E. (2002) In silico simulations reveal that replicators with limited dispersal evolve towards higher efficiency and fidelity. *Nature*, **420**, 340–343.
69. Hazen,R.M., Griffin,P.L., Carothers,J.M. and Szostak,J.W. (2007) Functional information and the emergence of biocomplexity. *Proc. Natl Acad. Sci. USA*, **104(Suppl. 1)**, 8574–8581.
70. Ancel,L.W. and Fontana,W. (2000) Plasticity, evolvability, and modularity in RNA. *J. Exp. Zool.*, **288**, 242–283.
71. Obermayer,B., Krammer,H., Braun,D. and Gerland,U. (2011) Emergence of information transmission in a prebiotic RNA reactor. *Phys. Rev. Lett.*, **107**, 018101.
72. Wu,M. and Higgs,P.G. (2011) Comparison of the roles of nucleotide synthesis, polymerization, and recombination in the origin of autocatalytic sets of RNAs. *Astrobiology*, **11**, 895–906.

73. Klussmann,S. (2006) *The Aptamer Handbook*. Wiley-VCH, Weinheim.
74. Turk,R.M., Chumachenko,N.V. and Yarus,M. (2010) Multiple translational products from a five-nucleotide ribozyme. *Proc. Natl Acad. Sci. USA*, **107**, 4585–4589.
75. Calderone,C.T. and Liu,D.R. (2004) Nucleic-acid-templated synthesis as a model system for ancient translation. *Curr. Opin. Chem. Biol.*, **8**, 645–653.
76. Ellington,A.D. (2009) Back to the future of nucleic acid self-amplification. *Nat. Chem. Biol.*, **5**, 200–201.
77. Kozlov,I.A., De Bouvere,B., Van Aerschot,A., Herdewijn,P. and Orgel,L.E. (1999) Efficient transfer of information from hexitol nucleic acids to RNA during nonenzymatic oligomerization. *J. Am. Chem. Soc.*, **121**, 5856–5859.
78. Lincoln,T.A. and Joyce,G.F. (2009) Self-sustained replication of an RNA enzyme. *Science*, **323**, 1229–1232.
79. Lohrmann,R., Bridson,P.K. and Orgel,L.E. (1980) Efficient metal-ion catalyzed template-directed oligonucleotide synthesis. *Science*, **208**, 1464–1465.
80. Luther,A., Brandsch,R. and von Kiedrowski,G. (1998) Surface-promoted replication and exponential amplification of DNA analogues. *Nature*, **396**, 245–248.
81. Mandel,J. (1982) Use of the singular value decomposition in regression analysis. *Am. Stat.*, **36**, 15–24.
82. Rohatgi,R., Bartel,D.P. and Szostak,J.W. (1996) Kinetic and mechanistic analysis of nonenzymatic, template-directed oligoribonucleotide ligation. *J. Am. Chem. Soc.*, **118**, 3332–3339.
83. R Development Core Team. (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
84. Usher,D.A. and McHale,A.H. (1976) Nonenzymatic joining of oligoadenylates on a polyuridylic acid template. *Science*, **192**, 53–54.
85. von Kiedrowski,G. (1986) A self-replicating hexadeoxynucleotide. *Angew. Chem. Int. Ed. Engl.*, **25**, 932–935.
86. Wehrens,R. and Mevik,B.-H. (2007) pls: partial least squares regression (PLSR) and principal component regression (PCR). R package version 2.1-0, http://mevik.net/work/software/pls.html (17 March 2010, date last accessed).