



Emergence of Information Transmission in a Prebiotic RNA Reactor

Benedikt Obermayer,^{1,*} Hubert Krammer,² Dieter Braun,² and Ulrich Gerland^{1,†}

¹*Arnold-Sommerfeld-Center für Theoretische Physik and Center for NanoScience, Ludwig-Maximilians-Universität München, Germany*

²*Systems Biophysics, Physics Department, Center for Nanoscience, Ludwig-Maximilians-Universität München, Germany*
(Received 18 March 2011; published 27 June 2011)

A poorly understood step in the transition from a chemical to a biological world is the emergence of self-replicating molecular systems. We study how a precursor for such a replicator might arise in a hydrothermal RNA reactor, which accumulates longer sequences from unbiased monomer influx and random ligation. In the reactor, intra- and intermolecular base pairing locally protects from random cleavage. By analyzing stochastic simulations, we find temporal sequence correlations that constitute a signature of information transmission, weaker but of the same form as in a true replicator.

DOI: 10.1103/PhysRevLett.107.018101

PACS numbers: 87.14.G–, 82.39.Pj, 87.15.H–, 87.23.Kg

The RNA world theory [1] posits that the first information carrying and catalytically active molecules at the origin of life were RNA-like polynucleotides [2]. This idea is empirically supported by the discovery of ribozymes, which perform many different reactions [3], among them the basic template-directed ligation and polymerization steps [4,5] necessary for replicating RNA. However, a concrete scenario of how a self-replicating RNA system could have arisen spontaneously from a pool of random polynucleotides is still lacking. Physical effects may have facilitated this step, as is believed to be the case in other transitions of prebiotic evolution [6].

From the perspective of information, an RNA replicator transmits sequence information from molecule to molecule, such that the information survives even when the original carrier molecules are degraded, for instance due to hydrolytic cleavage [7]. Rephrased in these terms, the problem of spontaneous emergence of an RNA replicator [8,9] becomes a question of a path from a short term to a lasting sequence memory. This transition occurred either as a single unlikely step or as a more gradual, multistep transition. Here, we explore a scenario of the latter type, based only on simple physicochemical processes (see Fig. 1): (i) random ligation of RNA molecules, e.g., in a hydrothermal “RNA reactor,” where polynucleotides are accumulated by thermophoresis [10], (ii) folding and hybridization of RNA strands, and (iii) preferential cleavage of single- rather than double-stranded RNA segments [7]. Using extensive computer simulations and theoretical analysis, we study the behavior that emerges when these processes are combined.

Clearly, the preferential cleavage at unpaired bases effectively creates a selection pressure for base pairing in the reactor. We find that this effect increases the complexity of RNA structures in the sequence pool, which may favor the emergence of ribozymes. The underlying sequence bias also extends the expected lifetime of sequence motifs in the finite pool. Interestingly, we find that correlations between

motifs persist even longer than expected. This memory effect is associated with information transmission via hybridization. Intriguingly, these correlations have the same statistical signature as templated self-replication, only weaker. In this sense, the RNA reactor could constitute a stepping-stone from which a true RNA replicator could emerge, e.g., assisted by a primitive ribozyme catalyzing template-directed synthesis.

RNA reactor.—As illustrated in Fig. 1, we envisage an open reaction volume V under nonequilibrium conditions as, e.g., inside a hydrothermal pore system where polynucleotides are strongly accumulated by a combination of convective flow and thermophoresis [10]. At any point in time, the reaction volume contains various sequences S_L of length L . The full time evolution of this pool is a stochastic process with the reactions

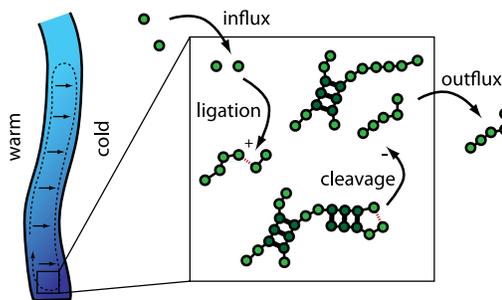


FIG. 1 (color online). Illustration of the RNA reactor. Left: Combined action of convection and thermophoresis in narrow pores subject to a temperature gradient results in strong accumulation of nucleotides, as indicated by the darker shading. Right: The region of high concentration defines an open reaction volume where nucleotides enter and bonds are formed through ligation reactions. Equilibrium base-pair formation protects bonds next to paired nucleotides (dark) from cleavage. Length-dependent outflux accounts for the preferential accumulation of long molecules.

$$\emptyset \xrightarrow{J} S_1, \quad S_L \xrightarrow{d_L} \emptyset, \quad (1a)$$

$$S_L + S_K \xrightarrow{\alpha} S_{L+K}, \quad S_L \xrightarrow{\beta_{L,K}} S_K + S_{L-K}. \quad (1b)$$

We assume a constant and unbiased influx of monomers (ACGU) at rate J . The effective outflux rate $d_L = d_0 e^{-(L/L_c)^{1/2}}$ accounts for the strong accumulation of nucleotides in a pore system, with a characteristic length dependence determined by the length scale L_c , which comprises parameters such as Soret coefficient, temperature gradient, and geometry [11]. Ligation of monomers or oligomers occurs at fixed rate α [12]. Finally, the most essential ingredient is a backbone cleavage process with a rate that depends on the base-pairing probability of the neighboring bases, such that double-stranded RNA is more stable than single-stranded RNA. Specifically, we calculate the cleavage rate $\beta_{L,K} = \beta_0(1 - p_{L,K})$ at backbone bond K using the average base-pairing probability $p_{L,K}$ of the two neighboring bases. We allow both intramolecular base pairs within single sequences and intermolecular base pairs within duplexes of any two molecules. RNA folding is performed by means of the Vienna package [15,16], where the partition function of the entire ensemble is calculated assuming chemical equilibrium [17], warranted by the fast hybridization kinetics [8].

We use the standard Gillespie algorithm to simulate the stochastic dynamics (1) of the sequence pool. The cleavage rate $\beta_{L,K}$, which is recalculated from the folding output for all molecules whenever necessary, effectively introduces a selection for base-pair formation. Since RNA folding depends on the temperature T and duplex formation is also concentration-dependent, we can vary the selection pressure via $p_{L,K}(T, V)$. We consider the reactions (1) under different possible conditions, with two different temperatures (a cold system at 10 °C and a hot environment at 60 °C) and concentrations (in the pM and mM range, respectively). To study the differences from a random pool, we also consider a “neutral” scenario without folding ($p_{L,K} = 0$). These scenarios are chosen mainly to highlight the effects of base pairing and not to suggest specific environmental conditions at the origin of life.

Stationary length and shape distribution.—Disregarding sequence-dependent selection, the ligation-cleavage dynamics of the RNA reactor resembles the kinetics of cluster aggregation and fragmentation. Hence, the stationary sequence length distribution shown in Fig. 2(a) corresponds to a cluster size distribution, and its moments can be obtained using established methods [16,18]. In the limit of large influx J , the average total molecule number $\langle N_{\text{tot}} \rangle$ and their mean length $\langle L \rangle$ are given by

$$\langle N_{\text{tot}} \rangle = \sqrt{\frac{J(d_0 + \beta_0)}{\alpha d_0}}, \quad \langle L \rangle = \sqrt{\frac{J\alpha}{d_0(\beta_0 + d_0)}}, \quad (2)$$

where we have neglected the length dependence of the outflux ($L_c \rightarrow \infty$; a finite value for L_c shifts both $\langle N_{\text{tot}} \rangle$ and $\langle L \rangle$ to larger values without strongly affecting the

shape of the distribution). These analytical results readily explain why with stronger selection the total number of molecules decreases, but their mean length goes up [see Figs. 2(b) and 2(c)]: the cleavage rate β_0 is reduced as the mean base-pairing probability $\langle \bar{p}_L \rangle$ is increased especially for longer sequences [cf. Fig. 2(d)], and the distribution thus gains more weight in the tail of long sequences.

In order to characterize the structural repertoire of this RNA pool, we focused on the tail of the length distribution and analyzed the secondary structures of long sequences with $L > L^*$. We performed the analysis for $L^* = 35$ as well as $L^* = 50$ (the length of the minimal hairpin ribozyme [19]). Figure 2(f) shows the probability to observe structures within basic “shape” classes [20], such as hairpins or hammerheads [21]. We observe a significant enrichment of complex structures under selection compared to the neutral case defined above.

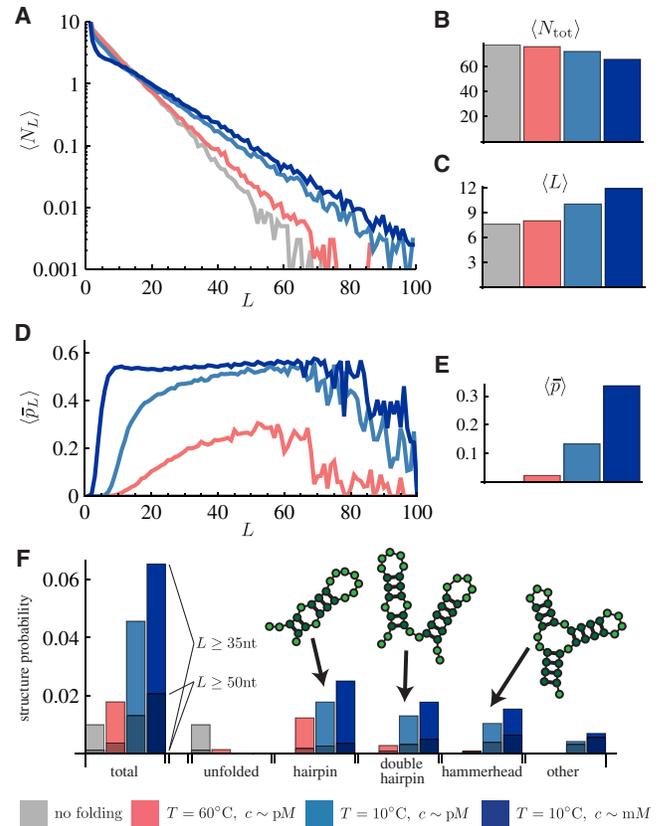


FIG. 2 (color online). Steady-state properties of the sequence pool: (a) length distribution $\langle N_L \rangle$, (b) total number $\langle N_{\text{tot}} \rangle$ of molecules, and (c) their mean length $\langle L \rangle$. (d) Base-pairing probability $\langle \bar{p}_L \rangle$ averaged over sequences of length L , with mean $\langle \bar{p} \rangle$ shown in (e). (f) Structural repertoire of long sequences: steady-state probabilities for sequences longer than $L^* = 35$ (shaded parts: $L^* = 50$), which fold into a structure of similar shape as the indicated schematic drawings. Selection strength increases from light to dark color as indicated in the legend. All observables are averaged over time and 10 independent replicas. Remaining parameter values were $J = 1$, $\alpha = 0.001$, $\beta_0 = 0.01$, $d_0 = 0.005$, $L_c = 10$.

Information transmission via hybridization.—Base pairing and the ensuing correlations between sequences occur mostly within relatively short sequence regions. Therefore, we focus on the dynamics of shorter subsequences or “sequence motifs” of length ℓ , which are informational entities not tied to a specific molecule. From our simulations, we extract time trajectories for the copy numbers $n_i(t)$ of all 4^ℓ different motifs. Even for fairly small $\ell > 3$, the sequence space of motifs is not fully covered in the finite ensemble; i.e., an average motif copy number is typically $\langle n_i(t) \rangle \ll 1$. Hence, signatures of information transmission should appear as an unexpected increase in the lifetime of these motifs. Suitably averaged observables are provided by the auto- and cross-correlation functions, $C_a(t) = 4^{-\ell} \sum_i \langle n_i(t) n_i(0) \rangle$ and $C_c(t) = 4^{-\ell} \sum_i \langle n_i(t) n_i^*(0) \rangle$, respectively, where n_i^* is the copy number of a motif’s (reverse) complement [21]. Figures 3(a) and 3(b) show data for these correlation functions for $\ell = 6$ and the parameter set used in Fig. 2.

The observed motif correlations can be understood in the framework of a simple stochastic process. Motifs are created when sequence ends are ligated together and destroyed by cleavage [22]. Using a mean-field-type approach, we pick an arbitrary probe motif with copy number $n(t)$. Its dynamics is described by a birth-death process, where $n(t)$ is increased with constant rate k_+ and decreased with linear rate k_- [see schema (i) in Fig. 3(c)]. The birth rate k_+ can be computed from the steady-state length distribution $\langle N_L \rangle$ by counting how many ends of long enough molecules are available for ligation. Assuming an annealed random ensemble, we obtain

$$k_+ = \frac{\alpha}{4^\ell} \sum_{k=1}^{\ell-1} \sum_{L \geq k} \langle N_L \rangle \sum_{L' \geq \ell-k} \langle N_{L'} \rangle. \quad (3)$$

The death rate k_- comprises the effects of cleavage and hybridization. A motif is cleaved with rate β_0 at any of its $\ell - 1$ bonds, but this rate is reduced by the effective base-pairing probability of its parent sequence, which in turn depends on the selection strength. On average, this reduction follows from averaging over the length and base-pairing probability distributions $\langle N_L \rangle$ and $\langle \bar{p}_L \rangle$ of parent sequences, respectively. This gives the result

$$k_- = \beta_0(\ell - 1) \left[1 - \frac{\sum_{L \geq \ell} (L - \ell + 1) \langle \bar{p}_L \rangle \langle N_L \rangle}{\sum_{L \geq \ell} (L - \ell + 1) \langle N_L \rangle} \right]. \quad (4)$$

However, a birth-death process based on these two effective rates alone necessarily fails to describe cross-correlations between a motif and its complement [23]. The reduction in the cleavage rate of a particular motif due to hybridization is conditional on the presence of its complementary partner. Hence, we modulate the average death rate k_- with an additional factor $h(x) \leq 1$, which accounts for the probability of hybridization and depends on the number $x = n^*/n$ of available complements per motif. Since the average hybridization probability is small under the conditions

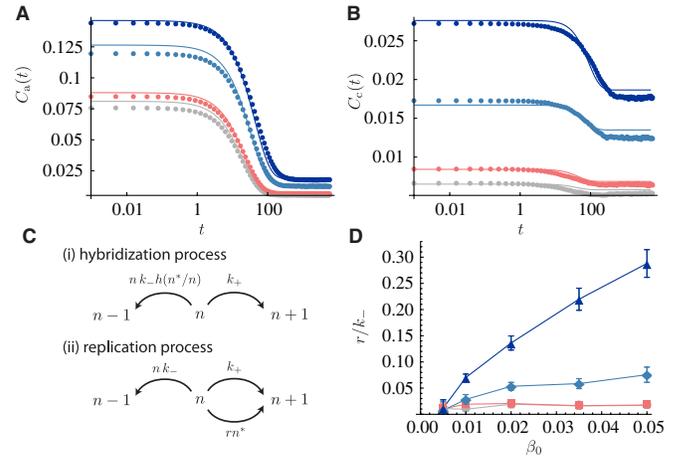


FIG. 3 (color online). Information transmission among sequence motifs. (a) and (b) Auto- and cross-correlation functions $C_{a,c}(t)$ from simulation data for $\ell = 6$ (dots) together with analytical expressions from Eq. (6) (solid lines). The rates k_- and k_+ have been computed from Eqs. (3) and (4), with r as the only fit parameter. (c) Schemata for different birth-death processes: (i) motifs are created with constant rate k_+ and destroyed with linear rate $k_- h(n^*/n)$, which is reduced by hybridization to their complements; (ii) motifs are destroyed with fixed rate k_- , but are copied from their complements with rate r . To leading order in r/k_- , both processes give rise to identical correlation functions $C_{a,c}(t)$, where a nonconstant $C_c(t)$ signifies information transmission between a motif and its complement. (d) Dependence of the replication efficiency r/k_- on the cleavage rate β_0 (error bars indicate 95% confidence intervals). Color code as in Fig. 2.

considered here, it will be proportional to x . This leads us to a linear ansatz $h(x) \approx 1 - (r/k_-)x$, where the significance of the coefficient r will shortly become apparent. We find that in the “hybridization process” of Fig. 3(c), the expected copy number $\langle n \rangle$ of a motif obeys

$$\partial_t \langle n \rangle = k_+ - k_- \langle n h(n^*/n) \rangle \approx k_+ - k_- \langle n \rangle + r \langle n^* \rangle. \quad (5)$$

A symmetric equation holds for $\langle n^* \rangle$. Strikingly, this result is identical to the corresponding rate equations for a “replication process” [16], where motifs are born with rate k_+ , destroyed with fixed rate k_- , and copied from their complements with rate r , as in schema (ii) of Fig. 3(c). This observation suggests that we may interpret the coefficient r as an apparent replication rate for motifs in the RNA reactor.

To validate this interpretation, and to measure the apparent replication rate in our simulations, we calculate the correlation functions of the hybridization process using the same approximation for $h(x)$ [16], yielding

$$C_{a,c}(t) = \frac{k_+^2}{(k_- - r)^2} + \frac{k_+ e^{-(k_- - r)t}}{2(k_- - r)} \pm \frac{k_+ e^{-(k_- + r)t}}{2(k_- + r)}. \quad (6)$$

In Figs. 3(a) and 3(b), we used these expressions with the rates k_+ and k_- calculated from Eqs. (3) and (4), and with r as the only free parameter fitted simultaneously to both data sets. The equivalence between the hybridization and

the replication processes is also exhibited by their correlation functions to leading order in r/k_- [16]. Hence, the good agreement with the simulation data indicates that the observed motif correlations are virtually indistinguishable from those expected for inefficient template-directed replication. The replication efficiency r/k_- determined by the fits is plotted in Fig. 3(d) as a function of the bare cleavage rate β_0 for the different conditions. Remarkably, it reaches levels close to 30% in the cold and highly concentrated environment, where base pairing via duplex formation is favorable. Note that a true (exponential) replicator would require that motifs are copied faster than they are degraded ($r > k_-$), while our system with $r < k_-$ is an inefficient realization.

These findings show that protection against cleavage due to folding and hybridization leads to an extended sequence memory in the RNA reactor. One global contribution to this longer motif lifetime is due to the “protection factor” in square brackets in Eq. (4), which renormalizes the bare cleavage rate to account for the average probability that a motif is paired. Another contribution stems from the correlation time in Eq. (6), which is increased as the apparent replication rate is subtracted from the renormalized cleavage rate, such that $C_{a,c}(t)$ decays on time scales of order $(k_- - r)^{-1}$. This specific increase occurs only when a motif and its complement mutually protect each other, and it therefore demonstrates the emergence of information transmission.

Conclusions.—We have analyzed stochastic simulations of a minimal prebiotic RNA reactor, where formation of double strands protects sequence parts from degradation. On the one hand, this selection for structure biases the resulting pool towards longer and more structured sequences, favoring the emergence of ribozymes. On the other hand, it leads to a weak apparent replication process based on “information transmission by hybridization,” conceptually similar to “sequencing-by-hybridization” techniques [24]. Together, the structural complexity and the information transmission featured in the RNA reactor suggests this type of system as plausible intermediate for the emergence of a true replicator with $r > k_-$. For instance, some of the relatively frequent simple structures observed in our simulation are similar to known ligase ribozymes [3]. This functionality in turn would facilitate the creation of more complex molecules from essential modular subunits [25]. Once ribozymes emerge, a self-replicating system could be established by template-directed ligation of suitably complementary oligomers [4]. So far, it has remained unclear how such autocatalytic RNA systems would be supplied with appropriate oligomer substrates. However, the strong cross-correlations observed in the RNA reactor demonstrate a significantly enhanced chance of finding sequences complementary to those present in the pool, including the sequence to be replicated. Thus, the RNA reactor acts as an adaptive filter to preferentially keep potentially useful substrate sequences. This adaptive selectivity would allow for the “heritable”

propagation of small variations and thus endow the replicator with basic evolutionary potential.

This work was supported by the Nanosystems Initiative Munich (NIM), by a DAAD grant to BO, and by a DFG grant to UG.

*Present address: Department of Physics, Harvard University, Cambridge MA 02138, USA.

†gerland@lmu.de

- [1] W. Gilbert, *Nature (London)* **319**, 618 (1986).
- [2] L. Orgel, *Crit. Rev. Biochem. Mol. Biol.* **39**, 99 (2004).
- [3] J. Doudna and T. Cech, *Nature (London)* **418**, 222 (2002).
- [4] N. Paul and G. F. Joyce, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12 733 (2002).
- [5] W. Johnston *et al.*, *Science* **292**, 1319 (2001).
- [6] I. Chen, R. Roberts, and J. Szostak, *Science* **305**, 1474 (2004).
- [7] D. Usher and A. Mchale, *Proc. Natl. Acad. Sci. U.S.A.* **73**, 1149 (1976).
- [8] C. Fernando, G. von Kiedrowski, and E. Szathmary, *J. Mol. Evol.* **64**, 572 (2007).
- [9] M. Nowak and H. Ohtsuki, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 14 924 (2008).
- [10] P. Baaske *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 9346 (2007).
- [11] S. Duhr and D. Braun, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19 678 (2006).
- [12] While nontemplated ligation occurs spontaneously [13] or via inorganic catalysis [14], we neglect template-directed reactions, which are less plausible in early prebiotic chemistry in the absence of ribozymes [2].
- [13] S. Pino, F. Ciciriello, G. Costanzo, and E. Di Mauro, *J. Biol. Chem.* **283**, 36 494 (2008).
- [14] J. Ferris and G. Ertem, *J. Am. Chem. Soc.* **115**, 12 270 (1993).
- [15] I. Hofacker *et al.*, *Monatsh. Chem.* **125**, 167 (1994).
- [16] See supplemental material at <http://link.aps.org/supplemental/10.1103/PhysRevLett.107.018101> for more details on the algorithm and the calculations, as well as supplementary results regarding GU pairs, self-complementarity, and shorter motifs.
- [17] S. H. Bernhart *et al.*, *Algorithms Mol. Biol.* **1**, 3 (2006).
- [18] R. Li, B. J. McCoy, and R. B. Diemer, *J. Colloid Interface Sci.* **291**, 375 (2005).
- [19] A. Hampel and R. Tritz, *Biochemistry* **28**, 4929 (1989).
- [20] R. Giegerich, B. Voss, and M. Rehmsmeier, *Nucleic Acids Res.* **32**, 4843 (2004).
- [21] Results shown in Figs. 2 and 3 were obtained disallowing ambiguous GU wobble pairs. See [16] for the length and shape distribution including GU pairs.
- [22] Since most motifs live on long sequences, we can neglect outflux reactions $\propto d_0 e^{-\sqrt{L}/L_c}$ against cleavage $\propto \beta_0 L$.
- [23] The presence of self-complementary sequences in a finite ensemble, which obey different statistics inherited by the corresponding motifs, leads to small cross-correlations even in the neutral case. See [16] for more details.
- [24] R. Drmanac *et al.*, *Advances in Biochemical Engineering/Biotechnology* **77**, 75 (2002).
- [25] C. Briones, M. Stich, and S. C. Manrubia, *RNA* **15**, 743 (2009).