# Cooperative Ligation Breaks Sequence Symmetry and Stabilizes Early Molecular Replication

Shoichi Toyabe[1,2] and Dieter Braun[1]

[1]Systems Biophysics, Physics Department, NanoSystems Initiative Munich and Center for Nanoscience, Ludwig-Maximilians-Universität München, Amalienstrasse 54, 80799 München, Germany
[2]Department of Applied Physics, Graduate School of Engineering, Tohoku University, Aramaki-aza Aoba 6-6-05, Aoba-ku, Sendai 980-8579, Japan

(Received 21 March 2018; revised manuscript received 11 February 2019; published 28 March 2019)

Each living species carries a complex DNA sequence that determines their unique features and functionalities. It is generally assumed that life started from a random pool of oligonucleotide sequences, generated by a prebiotic polymerization of nucleotides. The mechanism that initially facilitated the emergence of sequences that code for the function of the first species from such a random pool of sequences remains unknown. It is a central problem of the origin of life. An interesting option would be a self-selection mechanism by spontaneous symmetry breaking. Initial concentration fluctuations of specific sequence motifs would have been amplified and outcompeted less abundant sequences, enhancing the signal to noise to replicate and select functional sequences. Here, we demonstrate with experimental and theoretical findings that templated ligation would provide such a self-selection. In templated ligation, two adjacent single sequences strands are chemically joined when a third complementary strand sequence brings them in close proximity. This simple mechanism is a likely side product of a prebiotic polymerization chemistry once the strands reach the length to form double-stranded species. As shown here, the ligation gives rise to a nonlinear replication process by the cooperative ligation of matching sequences which self-promote their own elongation. This process leads to a cascade of enhanced template binding and faster ligation reactions. A requirement is the reshuffling of the strands by thermal cycling, enabled, for example, by microscale convection. By using a limited initial sequence space and performing long-term ligations, we find that complementary sequences with an initially higher concentration prevail over either noncomplementary or less-concentrated sequences. The latter die out by the molecular degradation that we simulate in the experiment by serial dilution. The experimental results are consistent with both explicit and abstract theory models that are generated considering the ligation rates determined experimentally. Previously, other nonlinear modes of replication such as hypercycles have been discussed to overcome instabilities from first-order replication dynamics such as the error catastrophe and the dominance of structurally simple but fast-replicating sequences, known as the Spiegelman problem. Assuming that templated ligation is driven by the same chemical mechanism that generates prebiotic polymerization of oligonucleotides, the mechanism could function as a missing link between polymerization and the self-stabilized replication, offering a pathway to the autonomous emergence of Darwinian evolution for the origin of life.

DOI: 10.1103/PhysRevX.9.011056

Subject Areas: Biological Physics, Interdisciplinary Physics, Nonlinear Dynamics

## I. INTRODUCTION

The genetic information of present-day living species is encoded in the DNA sequence as a specific combination of four different nucleotides: A, C, G, and T. How the first functional information coding sequences could have emerged from an initial pool of random sequences is one of the central questions to understand the origin of life.

Any sequence space of even a moderate length of, for example, 25 bases is so large ($4^{25} \approx 10^{15}$) that, even with a significant volume and concentration, the sampling can only be sparse, meaning that each molecule would have a different sequence. It must be expected that, even if such a very short sequence would have encoded and conferred an advantageous function for molecular evolution, it would not have had an impact on such a random pool of sequences. This limited sampling of sequence space is even the case for the most sophisticated systematic

evolution of ligands by exponential enrichment (SELEX) [1,2] lab procedures that are intended to select for functional molecules despite the fact that human brains and hands and complex machines are guiding the evolutionary process.

Spontaneous symmetry breaking is a basic physical mechanism that creates structure out of random initial conditions. We assess here whether nonlinear effects in a basic replication mechanism could implement the symmetry breaking in sequence space. Stochastic fluctuations in the initial sequence distribution would be amplified and give rise to an increasing homogeneous sequence space at a given location (Fig. 1). With the same process happening at different locations with stochastic variations, a large diversity of sequences could be sampled. Each sequence would be present at each location in significant concentrations such that a selection based on its function could be implemented. To test this scenario, we study how, under the replication dynamics of templated ligation, sets of similar sequences with high concentrations could survive by replication while less-concentrated or uncorrelated sequences die out.

It is no coincidence that a replication mechanism is the means to drive the self-selection dynamics of spontaneous symmetry breaking. Replication is necessary to later maintain and improve the functional sequence by the mechanism of Darwinian evolution. Therefore, the mechanism could be a route towards the emergence of sequence species for Darwinian evolution.

The understanding of the emergence of life has advanced and progressed fast in the past years. Examples include big steps forward in RNA catalyzed replication [3–10], synthesis of nucleotides [11,12], and base-by-base replication with activated nucleotides [13,14]. Autocatalytic replication of sequence information by the catalytic function of ribozymes is thought to be crucial for the evolution of central functions of biology. Autocatalytic replication has been demonstrated with carefully designed and selected ribozymes [15,16], where exponential growth of a group of mutually catalytic ribozymes is observed. However, the search of mechanisms allowing the spontaneous emergence of such complex autocatalytic sequences from a pool of random sequences remains an unsolved problem.

The first information molecules did not have many mechanisms at their disposal. After the synthesis and accumulation of the first nucleotides, random sequences [17–21] could polymerize. Once they are long enough to bind at the given temperature, three-molecule complexes form, and one sequence would bind to two complementary sequences which can then be connected by a suitable chemical reaction. This three-body reaction is termed a templated ligation. It offers a most basic replication mechanism, since the two sequences are linked only if the sequences match sufficiently, offering a transfer of information from one molecule to another (Fig. 1).

Our experiments reveal a reduction of the sequence space by replicating sequences with templated ligation. It is based on the fact that longer sequences are more likely to bind, and this replicates faster with the templated ligation. In addition, matching sequences can increase their length by ligation and therefore enhance their replication speed. As discussed with the experiments and theoretical descriptions below, this finding makes the replication nonlinear with respect to sequence concentrations.

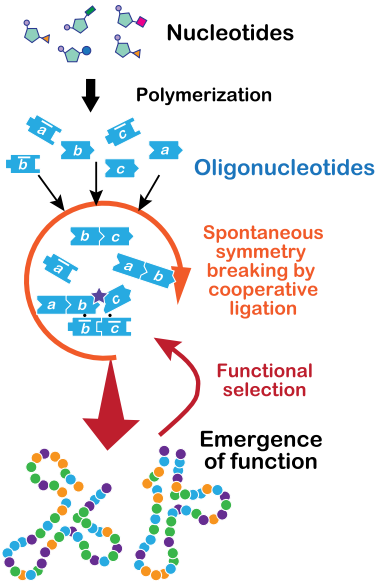


FIG. 1. Potential role of cooperative ligation in early molecular evolution of DNA oligonucleotides. Schematic representation of a potential path towards the selection and emergence of DNA sequences with incipient catalytic activities. Initially, a random oligonucleotide pool could have originated from the condensation and polymerization of single nucleotides. The same chemical mechanism that catalyzed the oligomerization of nucleotides could have also carried out the ligation of two adjacent oligonucleotides sequences brought into close proximity by a third complementary sequence. This process is termed templated ligation. The template binding would have required a critical oligonucleotide length at the temperature of polymerization. We use 20 mer oligomers at 67 °C to minimize the bias of the ligase protein used for ligation. In a pool of sequences, templated ligation becomes cooperative. Our experiments suggest that cooperative ligation could have led to a spontaneous breaking of sequence space symmetry. Depending on the initial concentration of ligated sequences at a given location, a cooperative set of sequences emerge that stabilize their own replication by ligation. Interestingly, the sequences with the highest initial concentration dominate the population, a process requiring a nonlinear replication characteristic. At a given location, this spontaneous symmetry breaking would generate a homogeneous oligonucleotide sequence pool. If this had an incipient function, the whole molecule population would show a significant evolutionary advantage, outcompeting the neighboring locations and creating a starting condition and stabilization for the emergence of Darwinian evolution.

In addition to the intractable search through a huge sequence space, early replicators faced two major instabilities: the error catastrophe [22] and a convergence towards

the fastest replicators, also known as the Spiegelman problem [23]. The first sequences are shaped by the balance between mutation and selection in a fitness landscape [22,24,25]. If the error rate exceeds a certain threshold, the selection can no longer suppress the accumulation of sequence errors and the sequences vanishes, which is termed an error catastrophe. This instability would have been a major bottleneck for early molecular evolution, since primitive replicators would initially have had only a limited fidelity. The dilemma is that strands require long and structured sequences in order to provide the necessary catalytic activity to decrease the error rate. But, for longer sequences, it is much harder to keep the error rate below the error threshold of the error catastrophe. Therefore, even if a self-replicating molecule did emerge at some point in a soup of random sequences, it would have been difficult for it to replicate only its own sequence information against the sequence majority of the pool.

The second dilemma is that first-order exponential growth tends to converge to a sequence with the highest replication rate. However, the faster replication of ever-shorter mutant sequences is very likely. Therefore, the inherent shortening of the strands by mutations likely suppresses the possibility to use oligonucleotides for the storage of sufficient information unless another process enhances the length of the sequences. In many experimental systems, the shortest possible sequence ends up dominating the population, as experimentally shown by Spiegelman [23,26] and also recently observed for the case of Ribo-PCR by Joyce [9]. Both the error catastrophe and the convergence to a common sequence would have strongly limited the emergence of early molecular Darwinian evolution.

The hypercycle proposed by Eigen and Schuster is a theoretical concept that can overcome both dilemmas [22]. A hypercycle is a ring-shaped network of replication reactions in which the product of a replication cycle catalyzes the reaction of the following replication cycle. This cooperative mechanism amplifies the sequence information at second order [20]; $\dot{x} = kx^2 – dx$. Here, $k$ and $d$ are the replication and the deletion rates, respectively, which can differ between replicators. Unlike the first-order growth, the growth rate of the hypercycle $\dot{x}/x = kx - d$ depends on $x$, and, thus, it is enhanced with the accumulation of $x$. Sexual reproduction is another example of higher-order growth [27]. That is, two elements react to generate the next generation of the elements. Such a system has been shown *in vitro* with an enzymatic DNA-RNA amplification system called cooperative amplification of templates by cross-hybridization (CATCH) [28,29]. The frequency-dependent selection, or the Allee effect [22], caused by the nonlinear growth of sequences could stabilize the wild types and could raise the error threshold. With the frequency-dependent selection, even replicators with smaller $k$ could survive and dominate if they once obtained a high frequency by fluctuations.

## II. RESULTS

### A. Templated ligation

The implementation of templated ligation experiments of DNA strands require, first, a ligation mechanism that could facilitate the formation of a phosphodiester bond between the 3'-hydroxyl and 5'-phosphate groups of two adjacent DNA strands and, second, the presence of a third template DNA strand that by complementary base pairing could bring the strands in close proximity for the ligation (Fig. S1 [30]).

Prebiotically plausible molecules that can carry out such a ligation reaction, in an efficient and rapid way, have not yet been identified. In this regard, diamidophosphate (DAP) is reported [31] as a potential prebiotic candidate molecule that could favor the oligomerization or condensation of nucleotides and amino acids into their respective polymers. However, the slow kinetic and low yield of such processes needs the support of physical nonequilibrium boundary conditions.

Faster methods and strategies exist, for example, the use of 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC) as *in situ* activator of phosphate groups of nucleotides, but the reaction suffers from modifications of the oligonucleotides at elevated temperatures. Furthermore, EDC in its optimized form is prebiotically not very plausible, but it can trigger both the polymerization and templated ligation of RNA or DNA.

Therefore, we decide to use a highly evolved protein to ligate the DNA strands. The ligation reaction experiments are conducted with a thermostable Taq DNA ligase [32] that catalyzes the ligation reaction >100-fold faster than EDC. We use elevated temperatures and relatively long sequences to reduce the impact of a sequence dependence in the catalytic activity of the ligase protein. In order to mathematically model the system, the sequence space is reduced to three different 20 mer DNA sequences denoted with lowercase letters, $\boldsymbol{a} = 5'$-atca gtgga agtgc tggtt, $\boldsymbol{b} = 5'$-atgag ggaca aggca acagt, and $\boldsymbol{c} = 5'$-attgg gtcac atcgg agtct and their reverse complements $\bar{\boldsymbol{a}}$, $\bar{\boldsymbol{b}}$, and $\bar{\boldsymbol{c}}$ with a single-base overhang at the 5' end to avoid blunt-end ligations. Capital letters denote both the respective sequences and their complements $A = \{\boldsymbol{a}, \bar{\boldsymbol{a}}\}$, $B = \{\boldsymbol{b}, \bar{\boldsymbol{b}}\}$, and $C = \{\boldsymbol{c}, \bar{\boldsymbol{c}}\}$. The sequences are designed to have comparable ligation rates, similar melting temperatures, and reduced self-annealing (Supplemental Material, Sec. S1 [30]). To allow the shuffling between hybridized strands prior to ligation, all experiments are conducted under the following thermal cycling conditions: 67 °C for 10 s and 95 °C for 5 s). Such thermal cycling could have been provided by thermal microscale convection [26,33] in prebiotic environments.

#### 1. Length dependence of templated ligation

Under the above-mentioned conditions, 20 mer sequences are found to bind less stably than 40 mers or 60 mer oligomers

to the same complementary 60 mer template sequence, and, therefore, longer sequences are elongated faster than shorter sequences [Figs. 2(a)–2(c)]. Template ligations are performed under thermal cycling with the template sequence $\overline{\boldsymbol{cba}}$. The kinetics of the ligation are measured for the following three sequence substrate sets: (a) $\boldsymbol{a}+\boldsymbol{b}$, (b) $\boldsymbol{a}+\boldsymbol{bc}$, and the competitive situation (c) $\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{bc}$ starting with the concentrations $[a]=100$ nM, $[b]=[bc]=33.3$ nM, and $[cba]=0.25$ nM. For the last case (c), the shorter strands $\boldsymbol{a}+\boldsymbol{b}$ compete with the longer strands $\boldsymbol{a}+\boldsymbol{bc}$ for ligation on the template. Under the competitive conditions (c), $\boldsymbol{a}+\boldsymbol{bc}$ ligate to form ***abc*** with a rate of 215 pM per cycle, 40-fold higher than the 5 pM per cycle rate calculated in order to obtain the elongation product from the shorter strands to form the sequence ***ab***. This competitive speed advantage of longer sequences leads to nonlinear replication.

Before going into the details of the cooperative ligation networks, we discuss how these ligation experiments are used to infer the experimental parameters of the subsequent models. As we see, the initial growth rates of ligation products provide both the dissociation constants $K_D$ of strand hybridization and the ligation rates $k$. To measure the initial growth rates $v_{ab}$, $v_{abc}$, $v'_{ab}$, and $v'_{abc}$ for $[ab]$ in (a), $[abc]$ in (b), $[ab]$ and $[abc]$ in (c), respectively, the product concentrations $[x](t)$ are fitted by a simple saturation curve

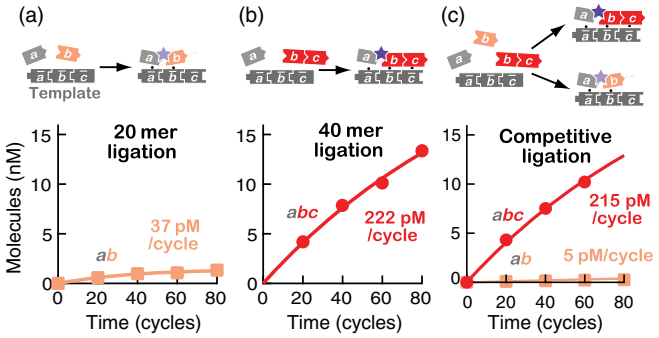


FIG. 2. Competitive ligation kinetics. To model the reaction, the kinetics of ligation is measured for (a) 20 mers, (b) 40 mers, and c) in the competitive case. Two DNA oligonucleotide sequences ***b*** and ***bc*** compete for ligation to a common complementary 60 mer template and ligate under thermal cycles between 67 °C for 10 s and 95 °C for 5 s without serial dilutions. The sequences of the 20 mers are $a$:5′-ATCAGGTGGAAGTGCTGGTT-3′, $b$:5′-ATGAGGGACAAGGCAACAGT-3′, and $c$:5′-ATTGGGTCACATCGGAGTCT-3′. Under competitive conditions, the templated ligation that gives rise to ***abc*** is 40 times faster than the ligation, leading to the ***ab*** sequence. Without competition, the difference reduces to a factor of 6. The kinetics could be understood by the ratio of the dissociation constants $K_D$ due to the competitive hybridization of ***b*** and ***bc*** on the same template sequence [Eq. (3)]. $K_D$ is modeled as an effective dissociation constant under thermal cycling. The $K_D$ for 20-base, 40-base, and 60-base nucleotide sequences could be determined by analyzing the ligation kinetics. The ligation rate $k$ is found to be constant in all three cases.

$[x](t)=(v_x/q)[1-\exp(-qt)]$ to find the initial growth rate $v_x$ at $t=0$ and the kinetics of saturation $q$ (Figs. 2 and S5 [30]). The saturation is caused by the inhibitory binding of the ligation products ***ab*** in (a) and (c) or ***abc*** in (b) and (c) to the template $\overline{\boldsymbol{cba}}$. As seen by the measurement results, the found saturation is minimal.

The thermal cycling is sufficiently fast such that the hybridization does not equilibrate during the cycles. Therefore, the experimental data are modeled by assuming that an effective dissociation constant $K_D$ defines the probability of binding under such fast thermal cycling conditions and that the hybridized strands are then ligated with a rate $k$. For the simplest case (a), the growth rate of the product ***ab*** is thus given by

$$\frac{d[ab]_0}{dt}=k_{ab}\frac{[a]_f[b]_f[\overline{cba}]_f}{K_{D,20}K_{D,20}}. \tag{1}$$

Here, $[\,]_f$ denotes the concentration of free, unhybridized strands, and $[\,]_0$ indicates the total concentration of strands. To infer the free-strand concentrations, the conservation laws for the binding of three strands are given by

$$\begin{aligned}
[a]_0&=[a]_f+\frac{[a]_f[\overline{cba}]_f}{K_{D,20}}+\frac{[a]_f[b]_f[\overline{cba}]_f}{K_{D,20}K_{D,20}},\\
[b]_0&=[b]_f+\frac{[b]_f[\overline{cba}]_f}{K_{D,20}}+\frac{[a]_f[b]_f[\overline{cba}]_f}{K_{D,20}K_{D,20}},\\
[\overline{cba}]_0&=[\overline{cba}]_f+\frac{[a]_f[b]_f[\overline{cba}]_f}{K_{D,20}K_{D,20}}+\frac{[a]_f[\overline{cba}]_f}{K_{D,20}}\\
&\quad+\frac{[b]_f[\overline{cba}]_f}{K_{D,20}}+\frac{[ab]_f[\overline{cba}]_f}{K_{D,40}}.
\end{aligned} \tag{2}$$

By solving these equations for the three cases ($a$, $b$, $c$) at $t=0$, we obtain

$$K_{D,20}=\alpha K_{D,40}\quad\text{and}\quad K_{D,40}=\frac{\beta-1}{\alpha-\beta}c_0, \tag{3}$$

where $\alpha=v_{abc}/v_{ab}$ and $\beta=v'_{abc}/v'_{ab}$ are obtained from the experiments.

The ligation rate $k$ is determined with the following reasoning. At small $t$, $[ab]_f=0$, $[a]_f\approx[a]_0$, and $[b]_f\approx[b]_0$, because the template concentration $[\overline{cba}]_0=0.25$ nM is considerably smaller than $[a]_0=100$ nM and $[b]_0=33$ nM. By inserting these values, we obtain the reaction rates for the ligation of ***ab*** from $\boldsymbol{a}+\boldsymbol{b}$ and that of ***abc*** from $\boldsymbol{a}+\boldsymbol{bc}$:

$$k_{ab} \simeq \frac{v_{ab}}{[\overline{cba}]_0}\left(1+\frac{K_{D,20}}{[a]_0}+\frac{K_{D,20}}{[b]_0}\right) \quad \text{and}$$
$$k_{abc} \simeq \frac{v_{abc}}{[\overline{cba}]_0}\left(1+\frac{K_{D,20}}{[a]_0}+\frac{K_{D,40}}{[bc]_0}\right). \tag{4}$$

This analysis results in the effective dissociation constants for the short and the long strands with $K_{D,20} = 193$ nM and $K_{D,40} = 4.5$ nM. The ligation rate instead is found to be $k_{ab} = 3.0\ \text{nM}^{-1}\,\text{cycle}^{-1}$ and $k_{abc} = 3.0\ \text{nM}^{-1}\,\text{cycle}^{-1}$. Since the ligation rate is therefore independent of the oligonucleotide strand length, we assume $k$ to be constant for all the sequences including substrates and templates. For case (b) and the competitive case (c) shown in Fig. 2, the binding of the ligated product ***abc*** on the template results in the saturation of the growth curves. By fitting the simulated curves with the parameters obtained above, we estimate $K_{D,60}$ to be 2.7 nM. In the subsequent simulations, we modelled the ligations with the effective dissociation constants estimated above $K_{D,20} = 193$ nM, $K_{D,40} = 4.5$ nM, and $K_{D,60} = 2.7$ nM for 20-base, 40-base and 60-base oligomers, respectively, under the applied temperature cycling.

### B. Ligation chain reaction under thermal cycling

So far, the elongation sequence product concentration is observed to increase linearly, since the complementary short strands are not added to the ligation reactions. The exponential product formation becomes important when, for instance, the pressure of a serial dilution, simulating the degradation of strands, exponentially removes molecules by, for example, the diffusion of molecules from the system. We therefore decide to implement a ligation chain reaction with exponential product formation by providing to the ligation reactions ***a***, ***b***, and ***c***, and also their complementary sequences $\bar{\boldsymbol{a}}$, $\bar{\boldsymbol{b}}$, and $\bar{\boldsymbol{c}}$, as described in Fig. 3(a). In this case, binding is not expected to be inhibited by the exponential growth due to the periodical temperature cycling and the sufficient supply of the above-mentioned shorter strands.

Importantly, if the ligation reaction does not reach an exponential phase, the replicates are gradually removed by serial dilution and die out. The molecule degradation present in early molecular evolution is in our experiments approximated by serial dilution. This worst-case approach avoids the possible complexities that could arise from the recycling of degraded sequences.

### C. Cooperative ligation network

Multiple ligase chain reactions, sharing overlapping sequences, could generate a cooperative ligation network. In the example shown in Fig. 3(b), the 40-base sequences depicted in the gray circle, $AB = \{\boldsymbol{ab}, \overline{\boldsymbol{ba}}\}$, $CA = \{\boldsymbol{ca}, \overline{\boldsymbol{ac}}\}$, and $BC = \{\boldsymbol{bc}, \overline{\boldsymbol{cb}}\}$, when supplied with the short 20-base

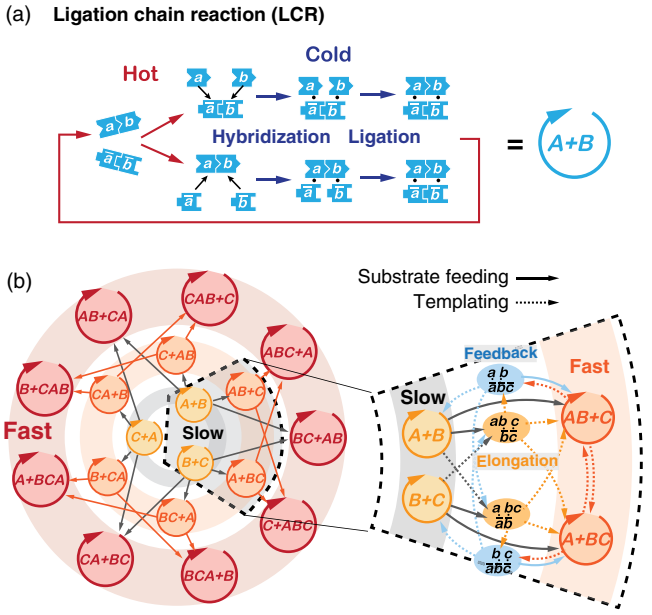


FIG. 3. Cooperative ligation chain reaction. (a) Schematic diagram of a ligation chain reaction depicting the exponential replication of oligonucleotide sequences under thermal cycling. Individual lowercase letters correspond to their 20-base single-stranded DNA sequences. (b) Schematic diagram of a cooperative ligation cascade network. Templated ligations of partially complementary sequences form a cooperative information-replication network. They are connected by binding to templates (dashed arrows) and by providing substrates (solid arrows). As seen in the enlarged section, the network expands towards longer sequences by ligation-based elongation. Longer sequences have more possibilities to work as templates and substrates and, therefore, mediate more reactions. Furthermore, the higher stability of longer strands (40-base and 60-base vs 20-base sequences) leads to faster ligation. This process creates a nonlinear cascade of cooperative replication reactions with a multistable dynamics. The replication became frequency dependent and sequences with higher initial concentrations are preferred, a characteristic not possible with the simple ligation chain reaction in (a).

sequence substrates $A = \{\boldsymbol{a}, \bar{\boldsymbol{a}}\}$, $B = \{\boldsymbol{b}, \bar{\boldsymbol{b}}\}$, and $C = \{\boldsymbol{c}, \bar{\boldsymbol{c}}\}$ can initially experience an exponential but slow growth, due to their short length and low stability to bind the substrates fragments with $K_{D,20}$. However, the sequences $AB$ and $BC$ can cooperate by complementary hybridizing at the common sequence $B$ and elongate, giving rise to the novel 60-base (yellow circle). Since the longer $ABC$ sequence can now act as a template for longer overlapping sequences, the ligations $AB + C \rightarrow ABC$ and $A + BC \rightarrow ABC$ show a faster ligation kinetics, since $K_{D,40} \ll K_{D,20}$, and provide significantly stronger binding. For this enhanced ligation reaction, the 40-base sequence substrates are provided by feeding from the slower ligation chain reactions with two letters (black arrows). But, in addition, the creation of the three-letter 60 mer sequences also offers an enhanced template to create $AB$ and $BC$ from $A$, $B$, and $C$ (blue circle). This combination of feed forward to

elongate the sequence (yellow) and feed backward to enhance the creation of substrates (blue) increases the concentrations of 40-base and 60-base sequences for cooperating sequences which are able to elongate. As documented with the following experiments, the increase in both complexity and overall ligation kinetics of the growing network therefore increases its replication speed as it progresses outward and gives rise to a concentration-dependent, higher-order replication dynamics.

### D. Simple cooperative ligation

To test a simple cooperative ligation network, we start the reaction with either a pair of two 40-base sequences that could ($AB$ and $BC$) or that could not cooperate ($AB$ and $AC$) in an elongation step with a common sequence pattern. The resulting elongated products are measured with gel electrophoresis [Figs. 4(a) and S4 [30] ].

For the noncooperative sequences $AB$ and $AC$ [Fig. 4(a), left], the logarithmic plot shows an initially exponential growth of the sequences $AB$ and $AC$. In comparison, the cooperative sequence pair $AB$ and $BC$ [Fig. 4(a), right] grows similarly in the first 40 cycles, but then the 60-base sequence ($ABC$) is found to grow at a faster rate. The longer sequence $ABC$ could act as an efficient template for the enhanced ligation rate of producing more $AB$ and $BC$ templates ($A + BC \rightarrow ABC$; $AB + C \rightarrow ABC$). This finding is confirmed by theoretical modeling as indicated below (Fig. 4, solid lines).

The concentrations of 40-base, two-letter sequence motifs, denoted by two capital letters in brackets ($\langle AA\rangle, \langle AB\rangle, \langle AC\rangle, \ldots$) are determined by COLD PCR. For example, the concentration $\langle AB\rangle$ indicates the concentrations of all motifs $AB$ in all present sequences. In Fig. 4, for example, $\langle AB\rangle$ corresponds to the sum of concentrations of the sequences $ab$, $abc$, $ba$, and $cba$. The measurement of the concentrations of the two-letter sequence motifs is possible experimentally by applying a quantitative PCR method with an initial low denaturation temperature (COLD PCR; further details are provided in Supplemental Material, Sec. S2 [30]) [29] which amplifies only 40 mers, even from longer strands. Note that deep sequencing would not have provided a comparable dynamic concentration range in terms of concentration to record motif concentrations between 0.1 pM and 100 nM needed to measure the sequence dynamics in our experiments. The method is calibrated with known 40 mer sequences.

When applied to the experiments in Fig. 4(a), we find the positive feedback by the elongating sequences. The motifs $\langle AB\rangle$ and $\langle BC\rangle$ grow about twofold more efficiently for the cooperative starting sequences as compared to the noncooperative pair $\langle AB\rangle$ and $\langle AC\rangle$ [Fig. 4(b)]. The cooperative ligation provides a significant boost in the presence of the common sequence motif $\langle AB\rangle$ and $\langle BC\rangle$ of the cooperating strands and provides more template concentration to grow more cooperating 40 mers from the supplied 20 mer sequences.

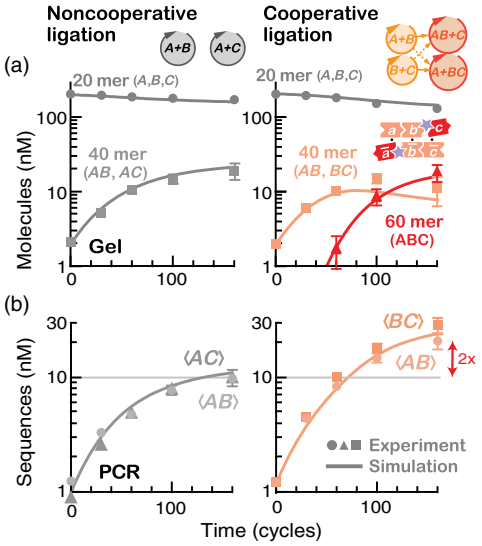


FIG. 4. Enhanced replication by cooperative ligation. (a) The ligation dynamics of the noncooperative sequences ($AB$ and $AC$) is compared with the cooperative sequences ($AB$ and $BC$). The latter could hybridize at the sequence $B$, creating a 60-base sequence template $ABC$ for faster ligations of $AB$ and $BC$. All reactions are supplied with the substrates sequences $A = \{\boldsymbol{a}, \bar{\boldsymbol{a}}\}$, $B = \{\boldsymbol{b}, \bar{\boldsymbol{b}}\}$, and $C = \{\boldsymbol{c}, \bar{\boldsymbol{c}}\}$. The concentrations of the oligonucleotides obtained after the ligation experiment are measured by gel electrophoresis as explained in Supplemental Material, Sec. S4 [30]. Solid lines are obtained from simulations [Eq. (9)] (Supplemental Material, Sec. S5 [30]) that considers the experimentally determined $K_D$ and $k$ parameters. (b) The concentration of the two-letter 40 mer sequence motifs $\langle NN\rangle$ is quantified by a calibrated real-time COLD PCR (Supplemental Material, Sec. S2 [30]). The cooperative ligation leads to a twofold faster growth rate of the 40-base sequence two-letter sequence motifs $\langle AB\rangle$ and $\langle BC\rangle$ compared to the growth rate of the noncooperative sequences $\langle AB\rangle$ and $\langle AC\rangle$. For example, $\langle AB\rangle$ combines the concentrations of the two sequences $[AB]+[ABC]$. Replicates of the experiment are shown in Fig. S10 [30].

### E. Cooperative ligation model

The deterministic rate equations are generated automatically with a program on Visual C# source code. Subsequently, the experiments are modeled on *Mathematica 10.1* (Wolfram research, IL, USA) with the obtained deterministic rate equations. The equations use the kinetic rates determined by the experiments in Fig. 2. Moreover, the equations take into consideration the binding by hybridization between strands. Conservation laws are used to infer the unbound molecule concentrations and to consider the serial dilutions to feed the substrates and dilute the products. Because of the inherent symmetry in the systems using both ($\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$) and their complementary sequences ($\bar{\boldsymbol{a}}$, $\bar{\boldsymbol{b}}$, $\bar{\boldsymbol{c}}$) in equal amounts, the rate equations are formulated in a simplified manner considering the

concentrations of $A$, $B$, and $C$. An overview of the modeling method is given below. Further details of the modeling are provided in Supplemental Material, Sec. S5 [30].

As already discussed, thermal cycling is not modeled explicitly, but we consider the experimentally determined ligation rate $k$ and effective dissociation constants $K_D$ (Fig. 2). In Figs. 4 and 5, because the strands do not elongate longer than 60 bases due to the limited number of initial 40-base template strands, we simulate all possible sequences. For the other experiments, it is checked that we can limit the maximum strand length to 120 bases for Fig. 5(a), to 640 bases for Fig. 6(b), and to 80 bases for Fig. 7 without affecting the simulation results significantly. We consider only binding complexes with up to three strands (Figs. 6 and 7) or four strands (Figs. 4 and 5). We estimate that the chance for higher-order complexes is much smaller than the three- or four-body complexes.

Despite the reduction of the complexity of the system and the use of a limited sequence space composed of only ***a***, ***b***, ***c*** sequences to study the dynamics, the resulting kinetic equation systems reach ASCII files for *Mathematica* with a size of around 30 MB. We include the Visual *C#* source code that is used to generate the equations in Supplemental Material [30]. In order to understand the approach, we show here the logic for the short reaction system of the experiment in Fig. 4.

As discussed, when measuring the ligation rates in Fig. 2, we determine the effective dissociation constants $K_{D,n}$ experimentally under thermal cycling with an $n$-base overlap. For an overlap longer than 60 bases, we use the same value as $K_{D,60}$. Qualitatively, this is a reasonable assumption based on the stronger binding thermodynamics but a slower binding kinetics due to the smaller concentrations of the binders. We use the following nomenclature to calculate concentrations of binding complexes on the assumption that the hybridization is effectively at equilibrium:

$$[A/AB] = \frac{[A]_f[AB]_f}{K_{D,20}},$$
$$[AB, C/ABC] = \frac{[AB]_f[C]_f[ABC]_f}{K_{D,40}K_{D,20}},$$
$$[AB, C/A, BC] = \frac{[AB]_f[C]_f[A]_f[BC]_f}{K_{D,20}K_{D,20}K_{D,20}}. \quad (5)$$

For example, $[AB, C/A, BC]$ denotes a four-body complex where $AB$ and $BC$ are hybridized with overhangs and $C$ and $A$ are hybridized to these overhangs. Some complexes have isomers; for example, $AA$ could hybridize to $AAC$ in two ways, one with only 20-base overlapping and one with a 40-base overlap and no overhang, both requiring different dissociation constants. Partially unbound complexes are included. For example, $ABC/ACC$ has a 40-base overlap in total; therefore, we use $K_{D,40}$ for this hybridization:

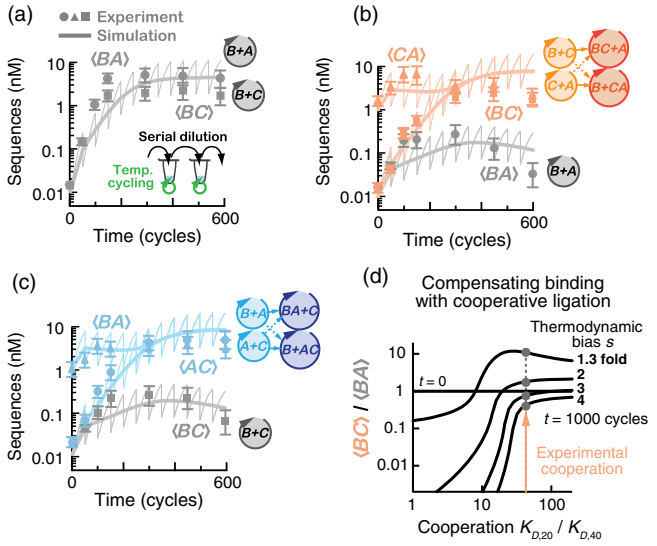


FIG. 5. Concentration-dependent replication by cooperative ligation. Cooperative ligation is found to replicate differentially based on the initial concentrations of the cooperating sequences. The reactions are performed under thermal cycling and serial dilution. The dilution simulates molecule degradation and is used to replenish ligase and a complete pool of 20-bases $A$, $B$, and $C$, simulating an untemplated polymerization as the oligonucleotide source. The three cases are as follows. (a) 40-base templates $BA$ and $BC$ are found to replicate from 0.01 nM (initial concentration) to similar steady state concentrations. 60-base sequences could not be formed due to the lack of a common binding motif. (b) The same ias performed with the addition of 1 nM of $CA$ sequence to the reaction. Since the sequences $CA$ and $BC$ could cooperate by forming the motif $BCA$, we observe that both the sequences survive the dilution effect. However, the noncooperating sequence $BA$ dies out. The initial concentration bias triggers a symmetry breaking by a concentration-dependent sequence selection. (c) Conversely, when the $BA$ sequence is provided to the reaction, the system picks $AC$ instead of $BC$ and creates the 60 mer sequence $BAC$. The thin lines correspond to mathematical simulations that consider the serial dilution dynamics. Thick curves connect the averages before and after the serial dilutions for comparing the simulation with experimental data. Replicates of the experiment are shown in Fig. S10 [30]. (d) We test how the kinetically driven cooperation could compensate a thermodynamic binding bias of the sequences $A$, $B$, and $C$. The reaction in (b) is simulated with a thermodynamic bias where sequences $A$ and $B$ bind $s$-fold better than the sequence $C$. Without a cooperative ligation mechanism ($K_{D,20}/K_{D,40} = 1$), the sequence $BA$ dominates after 1000 cycles over $BC$. With the experimental value $K_{D,20} = 43K_{D,40}$, the sequence $BC$ dominates over $BA$. The cooperative kinetics could overcome an up to $s = 3$-fold thermodynamic bias, offering an increased sequence diversity.

$$[ABC/ACC] = \frac{[ABC]_f[ACC]_f}{K_{D,40}},$$
$$[AB, C/BBC] = \frac{[AB]_f[C]_f[BBC]_f}{K_{D,20}K_{D,20}}. \quad (6)$$

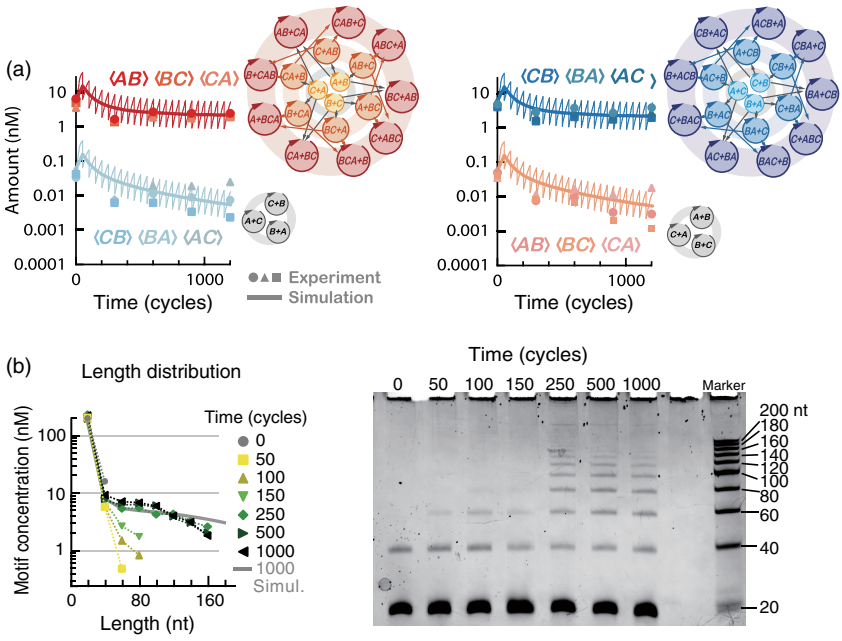


FIG. 6. Frequency-dependent selection of cooperative ligation sequences. Frequency-dependent replication of two competing groups of cooperating ligation patterns. Two groups of three cooperative sequences $AB$, $BC$, and $CA$ could generate the sequence "…$BCABC$…" by elongation, whereas the sequences $CB$, $BA$, and $AC$ could cooperate to generate the sequence "…$BACBA$…". (a) The reaction is started either with a concentration bias of 6 nM vs 60 pM for $AB$, $BC$, and $CA$ over $CB$, $BA$, and $AC$ or in the inverted concentration situation. As before, the replication reactions are supplemented with $A$, $B$, and $C$ sequences and subjected to a serial dilution that mimics a worst-case molecule degradation without sequence recycling. In both cases, the majority sequence motif patterns are sustained by replication against the serial dilution. The minority sequences die out. The simulation predicts the experimental finding. See Fig. S10 [30] for replicates of the experiment. (b) The cooperation generates long oligomers up to 160 bases in length from the initially 40-bases-long starting sequences after more than 500 cycles. The length distribution shows fat tails as measured by gel electrophoresis. The simulation predicts this length distribution (solid line).

The following conservation laws are added to the above hybridization expressions to obtain the total concentrations $[XY]$ on the left side from the free concentrations such as $[A]_f$ and $[AB]_f$. The occurrence of a given strand is collected from all the bound complexes. The conservation laws become, for example,

$$\begin{aligned}[A] &= [A]_f + [A/A] + [A/AB] + [A/CA] \\ &\quad + [A/BAC] + \cdots + [A, B/AB] + [A, BC/ABC] \\ &\quad + [C, A/CBA] + \cdots, \\ [AB] &= [AB]_f + [A/AB] + [AB/AB] + [A/CA] + \cdots \\ &\quad + [A, B/AB] + [AB, C/ABC] \\ &\quad + 2[AB, AB/ABAB] + \cdots. \end{aligned} \tag{7}$$

The ligation rate $k$ is not found to be dependent on the strand length (Fig. 2). Please note that ligations are modeled only if both end sequences of the substrates ($B$ and $C$ in this case) are matched with the template sequences and untemplated ligation is neglected. For example, we obtain

$$\begin{aligned}\dot{[A]} &= -k'(2[A, A/AA] + [A, B/AB] + [A, BC/ABC] \\ &\quad + [B, A/BA] + \cdots), \\ \dot{[AB]} &= k'([A, B/AB] + [A, B/CAB] + [A, B/ABC] \\ &\quad + [A, B/ABB] + \cdots) - k'([AB, C/ABC] \\ &\quad + [AB, C/AB] + [A, AB/AA] + \cdots), \\ \dot{[ABC]} &= k'([AB, C/ABC] + [AB, C/AB] \\ &\quad + [AB, C/BBC] + [A, BC/ABC] \\ &\quad + [A, BC/AB] + [A, BC/CAB] + \cdots) \\ &\quad - k'([ABC, C/CC] - [A, ABC/AAB] + \cdots). \end{aligned} \tag{8}$$

The reaction rate $k'$ is only slightly modified with $k'(t) = \alpha k f_n(t)$ from the experimentally determined rate of $k = 3.0\ \mathrm{nM}^{-1}\ \mathrm{cycle}^{-1}$. The coefficient $\alpha$ is necessary to model a saturation due to a limiting amount of the ligase. We use $\alpha = 1$ in Fig. 4, $\alpha = 1.27$ in Fig. 5, and $\alpha = 1.05$ in Fig. 6. We use larger $\alpha$ where the total concentrations of ligating complexes are low compared to that of Fig. 2. The term $f_n(t)$ models the degradation of the ligase by heating. We assume that the ligase degrades exponentially with a time constant of $\tau$; since fresh ligase is introduced by the

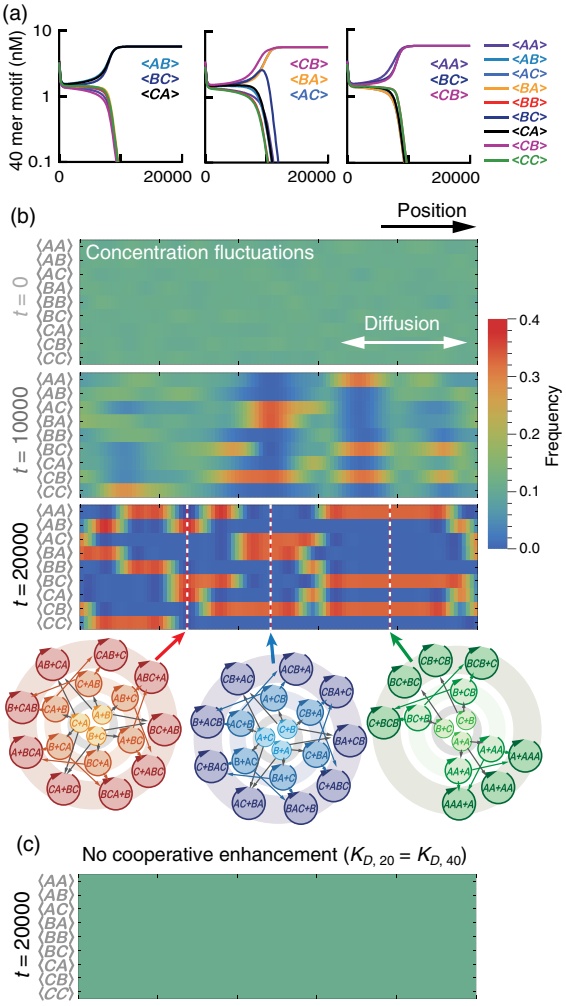


FIG. 7. Spontaneous symmetry breaking. (a) In a simulation of cooperative ligation, sequence structures emerge stochastically under well-mixed situations when we initiate the reactions by a 5% concentration fluctuation in an otherwise uniform concentration of all nine possible dimers ($AA, AB, \ldots, CC$). We observe a final steady state dominated by one of six possible cooperative combinations. For example, $AB$ and $AC$ do not coexist, because $AB$ and $AC$ are competitive as $B$ and $C$ compete for the right-hand side of $A$. The dilution rate is $d = 0.045$/cycle. (b) Spatial sequence patterns emerge and coexist despite the mixing by diffusion. Cooperating sequences emerge as local patterns after more than 10 000 cycles. Examples of the dominant ligation network are shown at the position of the broken lines. They correspond to the sequences in (a), which demonstrates the emergence of individual sequence patterns without the need of compartmentalization. (c) Control simulation without cooperative enhancement of the ligation from binding ($K_{D,20} = K_{D,40} = K_{D,60} = 150$ nM). The initial concentration fluctuations disappear, and the system converges to a uniform state.

serial dilution, in the $n$-th round of serial dilution, we assume $f_n(t) = (1-d_0)e^{-[t-(n-1)t_0]/\tau} + d_0 f_{n-1}(t)$ with $f_1(t) = e^{-t/\tau}$, where $d_0 = 1/6$ is the dilution ratio and $t_0 = 50$ cycles the period between serial dilutions. A degradation timescale of $\tau = 80$ cycles reproduces the experimental data well and could be expected even for a thermostable DNA ligase under our thermal cycling conditions. No additional modeling is needed to account for the binding of degraded ligase, suggesting that the degraded ligase does not actively block other ligation reactions.

The serial dilution in the simulation follows the experimental protocol: After 50 cycles, all strand concentrations are diluted by 1/6 except the 20 mer strands, which are also fed with $[X] \to 1/6[X] + 5/6X_0$, with $X_0 = 33$ nM the feeding concentration. Only for Figs. 7 and S8 [30] is a continuous degradation rate of $3.6\%/\text{cycle} = -(1/50)\ln(1/6)$ used for simplicity. To illustrate how the above building blocks converge into a system of rate equations, we show the simplest example, the reaction of Fig. 4 for the cooperative case of $AB$ and $BC$. The above rules lead to the rate equations (9). The effective ligation rate is $\hat{k} \equiv k\exp(-t/80)/K_{D,20}^2$ with an affinity ratio $p \equiv K_{D,20}/K_{D,40} - 1$:

$$\begin{aligned}
\dot{[A]} &= -\hat{k}([A][B]\{[AB]+[ABC]\} + [A][BC]\{[AB] \\
&\quad + (1+p)[ABC]\}), \\
\dot{[B]} &= -\hat{k}([A][B]\{[AB]+[ABC]\} + [B][C]\{[BC] \\
&\quad + [ABC]\}), \\
\dot{[C]} &= -\hat{k}([B][C]\{[BC]+[ABC]\} + [AB][C]\{[BC] \\
&\quad + (1+p)[ABC]\}), \\
\dot{[AB]} &= \hat{k}([A][B]\{[AB]+[ABC]\} - [AB][C][BC] \\
&\quad - (1+p)[AB][C][ABC]), \\
\dot{[BC]} &= \hat{k}([B][C]\{[BC]+[ABC]\} - [A][BC][AB] \\
&\quad - (1+p)[A][BC][ABC]), \\
\dot{[ABC]} &= \hat{k}\{[A][BC][AB] + [AB][C][BC] \\
&\quad + (1+p)[A][BC][ABC] + (1+p)[AB][C][ABC]\}.
\end{aligned} \tag{9}$$

### 2. *Long-term cooperative replication with serial dilution*

To test how the cooperative ligation affects the survival of sequences, we perform long-term replication experiments with various combinations of templates that are serially diluted. Every 50 cycles, 1/6 of the volume of the solution is diluted with a fresh solution containing 20-base DNA substrates and Taq DNA ligase. The dilution effect allows the simulation of an exponential long-term degradation of the template sequence and amounts to a simulated degradation of 3.6% per cycle against which the replication has to counteract to maintain the initial sequence pool of sequences.

First, the noncooperating template pairs $BA$ and $BC$ are investigated at an initial concentration of 0.01 nM [Fig. 5(a)]. Both withstand the exponential degradation, demonstrating their exponential replication, and settle into a steady state determined by the rates of replication and

serial dilution. This mutual symmetry breaks down as soon as the motif $CA$ is initially present [Fig. 5(b)]. The sequence pattern $\langle BA\rangle$ is suppressed and approaches extinction, but the sequence motif $\langle BC\rangle$ survives together with $\langle CA\rangle$. It has to be noted that the need for substrates is identical for both cases: All motifs compete with a second one for the substrates $A$, $B$, or $C$.

What breaks the symmetry in favor of $\langle BC\rangle$ and against $\langle BA\rangle$? Again, three-letter sequences emerge from two-letter templates. The metasequence $BCA$ assembles from $BC$ and $CA$ and helps both $\langle BC\rangle$ and $\langle CA\rangle$ in their replication [Fig. 4(b)] but offers no binding site for templated ligation with a cooperative partner for $\langle BA\rangle$. The alternative system where initially the template $AC$ is added instead of $CA$ inverts the preference and $\langle BC\rangle$ dies out. This result shows that the above asymmetry is not due to a thermodynamic bias of one particular 20 mer substrate or from a sequence dependent ligation rate, since now the cooperating 60 mer sequence $BAC$ emerges [Fig. 5(c)]. The simulation, based on the parameters found from Fig. 2, confirms the nonlinear growth and provides a quantitative description for the experiment [Figs. 5(a)–5(c), solid lines]. The length dependence of competitive ligation offers an enhanced replication of long consensus sequences and tips the balance towards the sequences which could collaborate by templated ligation with already existing motifs.

### 3. Kinetics of cooperation can overcome a thermodynamic bias

So far, sequences are chosen with similar hybridization strength for the same length of the oligonucleotide sequence. The question arises is how a concentration bias could outcompete over sequences which have a stronger hybridization and therefore a faster ligation kinetics in a frequency-dependent manner in Fig. 5. If this would not have been the case, stronger binding sequences would still invade even the majority of the population. To test this possibility, a sequence bias "$s$" is introduced to the ligation reactions by enhancing the thermodynamic binding stability of $A$ and $B$ compared to that of $C$. To keep the change symmetric, the $K_{D,20}$ for $A$ and $B$ is reduced to $K_{D,20}/\sqrt{s}$ and increases $K_{D,20}$ for $C$ to $K_{D,20}\sqrt{s}$ [Fig. 5(d)]. For consistency, we modify the dissociation constants of longer sequences and use $K_{D,40}/s$, $K_{D,40}$, $K_{D,40}$, and $K_{D,60}/\sqrt{s}$ for $BA$, $BC$, $CA$, and $BCA$, respectively.

We start a simulation with the initial concentrations of templates $BA$, $BC$, and $CA$ as in the experiment in Fig. 5(b). We plot the concentration ratio of $\langle BC\rangle$ over $\langle BA\rangle$ after 1000 temperature cycles versus the binding ratio of $K_{D,20}$ over $K_{D,40}$ which determines the cooperativity of the ligation network. The ratio indicates how much the initial concentration could determine the final replication outcome. For a vanishing length dependence of ligation, i.e., by switching off the advantage of cooperative ligation with $K_{D,20}/K_{D,40} = 1$, $\langle BA\rangle$ dominates over $\langle BC\rangle$ due to the introduced thermodynamic bias of binding. For example, for $s = 1.3$, as $K_{D,20}/K_{D,40}$ approaches 10, the cooperativity could overcome the thermodynamics bias. The motif $\langle BC\rangle$ now dominates over $\langle BA\rangle$ due to its faster ligation kinetics, a situation also formed at the experimental value of $K_{D,20}/K_{D,40} = 43$. Based on the model, cooperative ligation could therefore overcome a significant thermodynamic binding bias, which allows the system to amplify sequence motifs once they are present at an initially higher concentration even if they bind with lower affinity. The result would be an enhanced diversity of the accessible sequence space for evolution.

### 4. Frequency-dependent selection

In the previous experiments in Fig. 5, the length of the ligating sequences is limited to 60 mers due to the limited initial template set lacking templates that form overhangs necessary for an elongation. The same selection of the majority sequences is found for fully ligating templates. Now the sequences grow to considerable lengths [Fig. 6(b)]. We compare the replication of two competing groups of cooperating templates. On the one hand, the templates $AB$, $BC$, and $CA$ support the common, periodic motif $\ldots ABCABC\ldots$. When starting with templates $CB$, $BA$, and $AC$, the reverse motif $\ldots CBACBA\ldots$ would be expected to emerge. Those two motifs are two out of six possibilities of the two-letter sequences to cooperate (Fig. 7 and Supplemental Material, Sec. S6 [30]). Each of them is not promoting the ligation of the other two-letter sequences.

As seen in Fig. 6(a), the sequences which initially have a majority concentration establish and survive in a steady state, while the minority sequences decay exponentially as seen already in Fig. 5. For both opposing biases, we observe splitting of the growth kinetics and confirm that the initial concentration bias is amplified, even though the replication at lower concentrations is faster due to reduced saturation effects. The simulations of cooperative ligation predict the selection dynamics. In contrast, an exponential replicator without interactions would have immediately replicated all sequences to the identical, high concentration level as documented in Supplemental Material, Sec. S7 [30].

As seen in Fig. 6(b), the length distribution is initially exponential. But, after several cycles, long strands accumulate and form nonexponential fat tails. Similar shaped distributions were predicted by ligation models previously [18,34,35]. In our experiments, we find sequences with more than 160 bases, offering good support for the subsequent concentration-dependent selection of even more complex sequences.

### 5. Spontaneous symmetry breaking in simulated cooperative ligation networks

As we validate the theoretical model with the experiments, we use the model to extrapolate how individual

cooperating sequences emerge locally. Interestingly, the formed sequence patterns persist against the mixing mechanism of diffusion. Because of the nonlinear growth kinetics, the initial state with a near-uniform sequence concentration is found to be inherently unstable.

To extrapolate how the replication dynamics would amplify small concentration fluctuations, we perform a long-term simulation (Fig. 7). In addition to assuming a well-mixed situation [Fig. 7(a)], we implement molecular diffusion along one dimension. The experimental implementation of feeding, dilution flows, and diffusion in the physical space of such a setting is realized with microfluidics for the autocatalytic replication for the case of CATCH and $Q\beta$ replicase [29,36]. However, in the given setting, a similar stable long-term observation of ligation under thermal cycling and with the appropriate spatially resolved sequence analysis would be a highly challenging experiment.

The initial concentration of all two-letter template sequences $(AA, AB, AC, BA, \ldots)$ is superimposed with 5% random concentration fluctuations. Despite molecular diffusion, sequence patterns emerge after many temperature cycles. The combination of surviving two-letter sequences can be understood from the hierarchical replication structure. For example, $AB$, $BC$, and $CA$ cooperate towards sequences $\ldots BCABCA\ldots$ [Fig. 7(b), red], whereas $BC$, $CB$, and $AA$ converge to $\ldots BCBC\ldots$ and $\ldots AAAA\ldots$ [Fig. 7(b), yellow], representing two examples of six possible sequence patterns the system can converge to (Fig. S8 [30]).

The simulation parameters translate to a 5-mm-wide experimental system when using typical diffusion coefficients of the simulated oligomers [37] $D = 10^{-9}$ cm$^2/s$ and a temperature cycle of 30 s. The control simulation without cooperative rates of ligation ($K_{D,20}/K_{D,40} = 1$) produces no patterns [Fig. 7(c)]. Also, systems under well-mixed conditions but an initial 5% concentration fluctuation converge towards a randomly chosen single cooperative network where all other competing sequence networks die out (Supplemental Material, Sec. S6 [30]). This result demonstrates in a simulation that the nonlinear selection by the cooperative and hierarchical replication could amplify small fluctuations and spontaneously break the symmetry in sequence space.

## III. DISCUSSION

We argue with our experiments that cooperative interactions between complementary single-stranded DNA sequences under templated ligation create replication networks that implement two important properties of the hypercycle hypothesis [22,27,38]: (i) the achievement of a faster than exponential growth that results in a frequency-dependent replication, stabilizing the replication of past majorities sequences, and (ii) the inherent selection dynamics for the cooperating sequences in the ligation network. Both dynamic features are not due to special catalytic sequences at or near the ligation site [39,40] but caused by the simple fact that sequences that are elongated by ligation are able to bind stronger and therefore ligate faster. The mechanism requires only the minimal replication chemistry of ligation.

In the past, a number of theoretical explorations have indicated an interesting dynamic of ligation using coarse-grained modeling [33,41–44]. However, the lack of kinetic and thermodynamic details and the frequent inclusion of an experimentally not supported catalytic function prevent these models to yield quantitative experimental predictions for the presented cooperative ligation. The model by Tkachenko and Maslov indicates the possibility of nonlinear growth for templated ligation. To our knowledge, no experimental demonstration of nonlinear growth for cooperative ligation has been shown. Other modeling work on reaction networks [45–47] do not take into account sufficient mechanistic details to allow direct experimental implementation.

With the higher-order growth modes, the concentration of templates could enhance the speed of replication. The experiments show how majority sequence networks could suppress minority sequence networks. Growth becomes not only a function of the ligation rate. A network of cooperating ligation reactions could compensate for weaker binding [Fig. 5(d)]. The majority sequence could survive despite the suboptimal replication rate from inferior binding. The sequence history of replication becomes important: Once a cooperating majority sequence pattern emerges, it is inherently more stable and could defend itself favorably against emerging mutations, stabilizing the initially emerging sequences from statistical fluctuations. If we could include sequence degradation by hydrolysis into the reaction, we expect that the additionally provided sequences for templating would further stabilize the sequence majority.

This dynamic is seen in the experiment. High initial concentrations trigger a dominant ligation network that emerges from the cooperative hybridization dynamics between the sequences in the initial pool [48]. The nonlinear enhancement of growth by sequence matching is observed in the experiment already for the simplest binary cooperation (Fig. 4) but is also in trimeric cooperation networks [Figs. 5(b) and 5(c)] and for the formation of long sequences under long-term ligation [Fig. 6(a)]. The theoretical modeling supports the idea that the cooperativity ias caused by feedback and feed-forward mechanisms. For example, in Fig. 4, sequence matching 40 mer two-letter sequences concatenates into a feed-forward direction to form 60 mers with a three-letter sequence. In addition, a feedback loop back to two-letter sequences is created: The three-letter 60 mers template and replicate an increased concentration of two-letter sequences from single-letter sequences. The theoretical analysis finds that the

cooperation strength is given by the binding enhancement $p = K_{D,20}/K_{D,40} - 1$ [Eq. (9)]. Nonlinear growth and selection are observed only for $p > 0$ (Fig. 8). The value of $p$ is determined experimentally to about 40 (Fig. 2).

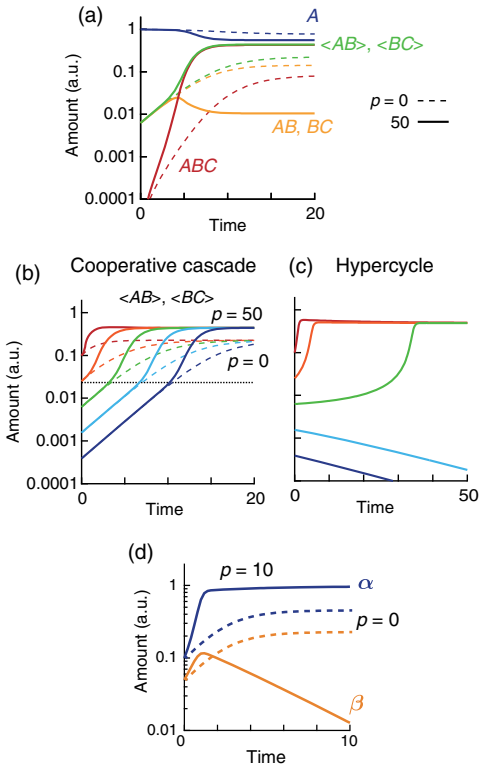


FIG. 8. Ligation cascade in simplified theoretical models. We analyze the theoretical model of our experimental system. A comparison with a hypercycle model illustrates the difference in their reaction orders and dynamics. (a) Cooperative ligation in Fig. 4 is simulated with a simplified model of Eq. (A1) with the parameters of effective reaction rate $\hat{k} = 1$, dilution rate $d = 0.6$, concentration of fed substrate $S = 1$, and a ligation enhancement by cooperation of $p = 50$ (solid lines) or $p = 0$ (dashed lines). The initial concentrations are $[A] = [B] = [C] = 1$. $[ABC] = 0$. $[AB] = [BC] = 0.1/16$. Growth curves of $[A]$ (navy), $[AB]$, $[BC]$ (orange), $[ABC]$ (red), and $\langle AB\rangle = [AB] + [ABC]$, $\langle BC\rangle = [BC] + [ABC]$ (green) with positive cooperativity ($p = 50$) are shown. (b) Growth curves start from different initial concentrations (0.1, 0.1/4, 0.1/16, 0.1/64, and 0.1/256) for the cooperative cascade model (A1) with $p = 50$ (solid lines) or $p = 0$ (dashed lines). The initial concentrations are changed to illustrate the difference in the growth dynamics. The growth is enhanced when the motif amount exceeds a threshold given by $\langle AB\rangle = \langle BC\rangle \approx 2[A]/p$ indicated by a horizontal dotted line. We use the steady state value of $[A]$. Hypercycles die out against dilution if started at a too low initial concentration. (c) Growth curves of the hypercycle model [15,43] with the same initial concentrations as in (b). (d) The concentration-dependent selection of cooperating sequences for the cooperative cascade model with $p = 10$ (solid lines) or 0 (dashed lines). Equations (B1) are numerically solved with the parameters $k = 1$, $S = 1$, $d = 0.5$, $\alpha(t = 0) = 0.1$, and $\beta(t = 0) = 0.05$. $L_\alpha$ is approximated by $\alpha$ for simplicity.

In the mechanism, the whole sequence determines the template binding for ligation. Specific bases at the ligation site are not essential for the mechanism. This contrasts with first-order base-by-base modes of replication [15,23], where sequences patterns with the highest thermodynamic stability have the tendency to dominate, leading to sequences biased towards G and C and presumably giving rise to a low sequence complexity. But molecules with a desirable catalytic function such as translation require sophisticated higher-order structures, which would be unlikely to develop from a purely thermodynamically dominated replication. In contrast, frequency-dependent selection from higher-order growth opens new routes for the emergence of complex functional molecules.

To study the basic properties of the cooperative cascade networks, we limit the number of unit blocks to three ($A$, $B$, and $C$). However, since the derivation of the nonlinear growth equations (12) does not explicitly depend on the number of unit blocks, the nonlinearity and the symmetry breaking shown in the result section should have also taken place with more unit blocks. We, indeed, expect that the use of more unit blocks would provide a richer selection dynamic. However, at the same time, the increase in the number of units would reduce the probability of binding between complementary DNA strands. While the probability to bind a complementary sequence would drop, it is not obvious why the concentration window for the onset of nonlinear growth would have changed with respect to the concentration regime of ligation.

Our proposed mechanism considerably differs from second-order growth models such as the hypercycle. In these models, the replication rate becomes proportional to the number of replicates, leading to hyperbolic growth [Eq. (13)]. Unfortunately, the hyperbolic growth also implies that for degradation conditions, mimicked in our experimental setting by serial dilution, the initial strand concentration must be kept over a concentration threshold. Otherwise, the replicators vanish due to the lack of a fast replication kinetics [Fig. 8(c)]. In contrast, the nonlinearity of a cooperative cascade of ligation offers a nonvanishing replication rate for low replicate concentrations. Even vanishing concentrations can be replicated in principle [Eq. (12) and Fig. 8(b)]. We do not clearly observe a downward convexity in the growth curves (Figs. 4–6). While the hypercycle has a strong downward convexity because of its pure second-order nature [Fig. 8(c)], the convexity is subtle in the present system because the first-order and second-order reactions are superposed [Fig. 8(b)]. A significant improvement in the quantification would be necessary to detect the subtle convexity.

It has been argued that classical hypercycles are sensitive to parasites that could take a free ride on the catalytic functions of the cycle [48,49]. The reason is that the cooperation is asymmetric and the product of one cycle acts as the replicating machinery for the next cycle. In the shown cooperative

ligation networks, however, the catalytic role played by the template is symmetric. If there is a reaction where a strand $X$ works as a template to produce a strand $Y$, there is also the reverse reaction where $Y$ acts as a template to produce $X$. Therefore, asymmetric free riders of one sequence which would not also offer themselves as a template would be difficult to imagine. The symmetric nature of the ligation network, in contrast to hypercycles, should therefore be much robust against simple sequence parasites.

In the past, well-balanced combinations of protein-catalyzed DNA reactions using multiple proteins have been shown to implement complex algorithms [28,50,51]. These experiments include higher-order growth dynamics. But these implementations were created from a fine-tuned network of complex protein functions which are hard, if not impossible, to imagine prebiotically. In contrast, we use a protein to speed up a prebiotically plausible ligation reaction. The mechanism we study does not require a particular catalytic enzyme, but the basic combination of hybridization and ligation. This reaction is likely implemented by the prebiotic activation chemistry that would have also triggered the random polymerization of the nucleotides in the first place.

Previously, replication mechanisms have been studied using the complex catalytic function of proteins to implement base-by-base replication such as $Q\beta$ replicase [23] or a combination of a reverse transcriptase and a polymerase [32]. These proteins generally show particular sequence dependencies and nontrivial reaction mechanisms such as template detection. The $Q\beta$ replicase under high salt concentrations could lead to an arbitrary elongation of sequences by untemplated bases [52]. However, under kinetic competition, as done in the classical experiments of Spiegelman [23], these replication mechanisms have the tendency to shorten the sequence length by mutations.

In comparison, the templated concatenation by ligation shows the inherent tendency to produce sequences longer than the initial templates. And since longer sequences offer more stable binding sites, they are able to cooperate with other sequences, and longer strands are favored in the ligation networks. We find a fat tail in the length distribution in our experiments [Fig. 6(b)] which enables significantly more complex sequences to emerge, enhancing the possibility of finding catalytically functional oligonucleotides. In the past, theoretical and experimental analysis of ligated sequences confirmed a reduction in the sequence diversity [18,34,35].

We measure the sequence dependence of the ligation kinetics and use it to model the ligation experiments (Fig. 2). We find that the ligation rate $k$ is constant and independent of the strand length once a bound template complex is established. However, the binding affinity $K_D$ determines the formation of the templated preligation complex and modulates the effective rate of ligation. Both kinetic rates are kept constant in the modeling.

Between experimental implementations, the rate $k$ requires only slight adaptations to fit the experiments, likely caused by saturating the catalytic activity of the protein due to varying concentrations of ligation sites. In addition, different batches of the ligase show slightly varied thermal degradation. Apart from these minor adaptations of the overall ligation characteristics, no fits of parameters are required to describe the experiments with the theoretical models, which is why we are confident to use the simulations to predict the system's behavior in scenarios which would have been experimentally too difficult to perform reproducibily.

In this study, we limit the number of unit building blocks to three pairs ($A$, $B$, and $C$) in order to study the basic properties of the cooperative cascade and still be able to follow the dynamics with the theory. However, there is no reason to assume the shown nonlinear replication that enables symmetry breaking would not take place, for example, for a more diverse sequence population or for shorter sequence lengths at lower temperatures. For example, the derivation of the nonlinear growth equations (12) does not depend on the number of different short sequences. We expect that the use of more starting sequences provides a richer dynamic. But, at the same time, the increase in sequence space would reduce the probability for a matching hybridization to happen, resulting in the need to enhance the concentrations of the starting strands in the experiments.

In addition to the numerical simulations, the analytical theory in the Appendixes is used to describe how the cooperativity causes the instability of the uniform state in the sequence space. The amplification from small concentration perturbations merges into patterns of majority networks and makes them stable in sequence space. The amplification of the fluctuations is predicted to produce different dominating sequence patterns at different locations even under the presence of diffusion. Once the spatial pattern forms, it would have been stabilized by the frequency-dependent selection, suppressing the growth of competing sequences diffusing from neighboring patterns. It is likely that in the long term only one sequence pattern remains. Before, the coexistence of multiple sequence configurations could only be realized without compartments such as membranes, limiting the potential for lateral gene transfer. It should be noted that the spatial pattern formation based on hypercycle, CATCH, and $Q\beta$ replicase characteristics has been explored previously by simulations [28,49] and in impressive microfluidic experiments [28,35,52]. An experimental test of the spatial coexistence of different sequence information under ligation and thermal cycles is one of the next experimental challenges.

In this study, ligation provides a primitive mode for the replication for sequences. As shown, the inherent cooperative characteristic of ligation would likely stabilize the replication of sequence information. It should be noted

that, for the full exploration of long sequences, the thermal cycling seems important to shuffle sequences between templates and to enable exponential replication. Limited versions of the dynamics should be, however, already seen for short sequences where the off rate is high enough to allow spontaneous exponential ligation chain reactions.

To implement temperature cycles, a heat flow, for example, across elongated rock pore systems on an early Earth [36,53], could provide thermal cycling by microfluidic convection [54,55]. Such a thermal convection could also provide the enhanced concentrations of oligonucleotides for templating and possibly increase their polymerization dynamics [20]. Combined with a flow-through, the thermal gradient could localize replicated oligonucleotides in a length-selective manner and supply the replication reactions continuously [26]. It should be noted that the shown mechanism bears a strong similarity to chemical systems that are able to amplify a bias between chiral molecules [56]. We therefore hypothesize that the concentration dependence of the ligation network could also purify backbone heterogeneity based on their differential duplex stability [57].

Ligation is a comparably simple reaction, and the chemical details of its prebiotic implementation have been explored [12,58–62]. It will be interesting to see how similar cooperative mechanisms could be established in base-by-base replication mechanisms. Replication of RNA or DNA from single bases using ribozymes has progressed continuously as shown by Joyce [9] and Holliger [63]. Noncatalytic replication is a very interesting third possibility, and recent experiments by Szostak show fast progress [64], becoming more attractive after better understanding its prebiotic plausibility [19]. In the future, it would be also interesting to integrate the network capabilities of RNA-based recombinases [8,16,65] to explore how the emergence of recombination as a most simple catalytic activity of RNA could implement modes of cooperative replication dynamics under thermal cycling.

## IV. CONCLUSION

Our experiments showed that a most basic mechanism—the joining of two DNA strands on a third template strand in a pool of diverse sequences—formed a cooperative cross-catalytic reaction network with higher-order replication kinetics. The shown process offers a mechanism that implements in a significantly more simple and robust manner the nonlinear replication dynamics first proposed by hypercycles. This nonlinear ligation chain reaction could circumvent central drawbacks of hypercycles, namely, the concentration growth threshold under molecule degradation [Fig. 8(c)] and the instability against "viral" molecules using the replication dynamics but not contributing to it.

It remains to be seen how the mechanism could amplify a majority sequence bias from more than the shown six sequences, i.e., from a more diverse random sequence pool. We expect that the hybridization dynamics of 20 mers would offer many more sequences to implement the shown mechanism. Unfortunately, it is challenging to extrapolate the brute-force theoretical modeling to more complex experiments. In the future, we expect long-term experiments using deep sequencing to offer more insights into this highly nonlinear, but realistic replication dynamics which was possible when the first oligonucleotides emerged.

## ACKNOWLEDGMENTS

This work was supported by the Alexander von Humboldt Foundation, European Research Council (ERC) Advanced Grant EvoTrap, the CRC Emergence of Life project P07, the SFB 1032 Project A04, a grant from the Simons Foundation (SCOL 327125, DB), and JSPS KAKENHI (15H05460). We very much thank Jonathan Liu, Patrick Kudella, Alexandra Kühnlein, and Noel Martin for corrections on the manuscript.

## APPENDIX A: NONLINEAR GROWTH MECHANISM

We analyze an analytical model to elaborate on the mechanism of the higher-order growth and the frequency-dependent selection. Instead of the huge set of equations used in the numerical simulations (Supplemental Material, Sec. S5 [30]), we use simplified rate equations. We consider a simple cooperative system starting with the templates $AB$ and $BC$. This system corresponds to the experiment shown in Fig. 3, but here we add dilution and feeding kinetics. The rate equations are

$$\begin{aligned}
\dot{[A]} &= -\hat{k}([A][B]\{[AB]+[ABC]\}+[A][BC]\{[AB] \\
&\quad +(1+p)[ABC]\}) - d([A]-S), \\
\dot{[B]} &= -\hat{k}([A][B]\{[AB]+[ABC]\}+[B][C]\{[BC] \\
&\quad +[ABC]\}) - d([B]-S), \\
\dot{[C]} &= -\hat{k}([B][C]\{[BC]+[ABC]\}+[AB][C]\{[BC] \\
&\quad +(1+p)[ABC]\}) - d([C]-S), \\
\dot{[AB]} &= \hat{k}([A][B]\{[AB]+[ABC]\}-[AB][C][BC] \\
&\quad -(1+p)[AB][C][ABC]) - d[AB], \\
\dot{[BC]} &= \hat{k}([B][C]\{[BC]+[ABC]\}-[A][BC][AB] \\
&\quad -(1+p)[A][BC][ABC]) - d[BC], \\
\dot{[ABC]} &= \hat{k}\{[A][BC][AB]+[AB][C][BC] \\
&\quad +(1+p)[A][BC][ABC]+(1+p)[AB][C][ABC]\} \\
&\quad -d[ABC].
\end{aligned} \tag{A1}$$

Here, $\hat{k} \equiv k/K_{D,20}^2$ and $p \equiv K_{D,20}/K_{D,40-1} - 1$. $S$ is the amount of the substrate supply, and $d$ is a degradation rate.

The growth curves obtained by simulating Eq. (A1) with $p = 50$ and $p = 0$ are plotted in Fig. 8(a). We see an enhanced growth of $ABC$ with $p = 50$ which eventually dominates the $\langle AB\rangle$ and $\langle BC\rangle$ motif concentrations. After $AB$ and $BC$ accumulate, $ABC$ grows due to the faster ligation rate (Fig. 2). This enhanced growth demonstrates the positive feedback mechanism by the cooperation of $AB$ and $BC$. $ABC$ mediates the cooperation between $AB$ and $BC$ and accelerates the growth of $\langle AB\rangle$ and $\langle BC\rangle$ with a downward convex. Without the enhanced ligation speed by a longer overlapping ($p = 0$), we do not observe the higher-order growth and find a lower steady state concentration of $\langle AB\rangle$ and $\langle BC\rangle$. These plots show that two factors account for the higher-order growth: the cooperation mediated by an elongation and the enhanced ligation rate for longer strands.

The growth curves of the cooperative cascade model [Fig. 8(b)] and the simplest hypercycle model [Fig. 8(c)] calculated starting from varied initial concentrations clearly show the difference of the two models. Both show a nonlinear growth, but with different characteristics. In the hypercycle model [Fig. 8(c)], the initial growth rate decreases as the initial concentrations decrease. A growth overcoming dilution is observed only with an initial concentration above some threshold. On the other hand, in the cooperative cascade model, the initial growth rates are similar for all initial concentrations. When the concentrations exceed a threshold, concentration indicated by a dotted horizontal line, the reaction is accelerated. The threshold is given by approximately $2[A]/p$, which is indicated by a dotted horizontal line in Fig. 8(b) as we see.

To understand these behaviors, we simplify the rate equations. By rearranging Eq. (A1), we obtain the rate equations of the sequence motifs $\langle AB\rangle \equiv [AB] + [ABC]$ and $\langle BC\rangle \equiv [BC] + [ABC]$:

$$\begin{aligned}\dot{\langle AB\rangle} &= k'[A]\{(S - \langle AB\rangle)\langle AB\rangle + p[BC][ABC]\} - d\langle AB\rangle,\\ \dot{\langle BC\rangle} &= k'[C]\{(S - \langle BC\rangle)\langle BC\rangle + p[AB][ABC]\} - d\langle BC\rangle.\end{aligned} \tag{A2}$$

By reasonably assuming the symmetry $[A] = [C]$ and $[AB] = [BC]$, we obtain the rate equation for $\alpha = \langle AB\rangle = \langle BC\rangle$ according to

$$\dot{\alpha} = k'(S - \alpha)^2(1 + pL_\alpha)\alpha - d\alpha. \tag{A3}$$

Here, $L_\alpha \equiv ([A]/[AB] + [A]/[ABC])^{-1}$ is the harmonic average of $[AB]/[A]$ and $[ABC]/[A]$ divided by 2, which is close to the minimum of either $[AB]/[A]$ or $[ABC]/[A]$. Because $L_\alpha$ quantifies the accumulation of long strands, the reaction rate $\tilde{k} = k'(S - \alpha)^2(1 + pL_\alpha)$ increases as the reactions proceed and the long strands accumulate. The growth is enhanced when the nonlinear term $pL_\alpha$ exceeds one. This threshold corresponds to $[ABC] \approx [A]/p$ or approximately $\alpha \approx 2[A]/p$ in the present setup.

The hypercycle model based on the template-directed ligations with an unknown hypothetical catalytic function by oligonucleotides is theoretically studied by Wills *et al.* [43]. They assume an unknown hypothetical catalytic ability by oligonucleotide to construct a hypercycle model. The rate equation of the hypercycle is reduced to

$$\dot{x} = k'(S - x)^2x^2 - dx. \tag{A4}$$

These simplified rate equations (A3) and (A4) explain the growth modes seen in Figs. 8(b) and 8(c). The hypercycle (A4) has a strong nonlinearity with the reaction rate proportional to $x$, which vanishes as $x$ goes to zero. Therefore, a growth that can overcome the serial dilution is not observed when started from low initial concentrations. On the other hand, in the cooperative ligation cascade model (A2), the reaction rate does not vanish even at a vanishing concentration $\alpha$ because of the moderate nonlinearity $(1 + pL_\alpha)\alpha$. Therefore, if the reaction rate $\tilde{k}$ is larger than dilution rate $d$, positive growth is always observed independent of the initial concentrations. The system grows in a manner that the reaction rate $\tilde{k}$ increases gradually from a slow rate $k'(S - \alpha)^2$ to a faster rate $k'(S - \alpha)^2(1 + pL_\alpha)$ as the concentration $\alpha$ increased.

## APPENDIX B: CONCENTRATION-DEPENDENT SELECTION OF SEQUENCE MOTIFS

We consider another cooperative sequence group $\beta$, which is, for example, $\langle BC\rangle$ and $\langle CA\rangle$ and shares the same substrates $A$, $B$, and $C$ with $\alpha$. The rate equation for the competitive growth of $\alpha$ and $\beta$ leads to

$$\begin{pmatrix}\dot{\alpha}\\ \dot{\beta}\end{pmatrix} = k'(S - \alpha - \beta)^2\begin{pmatrix}(1 + pL_\alpha)\alpha\\ (1 + pL_\beta)\beta\end{pmatrix} - d\begin{pmatrix}\alpha\\ \beta\end{pmatrix}. \tag{B1}$$

The selection dynamics simulated by Eq. (B1) is shown in Fig. 8(d). The frequency-dependent selection is observed only when $p > 0$, similar to the results in Fig. 6(a). This behavior can be explained by a linear-stability analysis of Eq. (B1). The stability of the uniform state, $\alpha_0 = \beta_0 \neq 0$, is determined by the sign of $p$ (see Supplemental Material, Sec. S7.2 [30]). When $p > 0$, the uniform state is unstable, and either $\alpha$ or $\beta$ becomes dominant. When $p = 0$, the uniform state is neutral. The initial difference is kept for a long time. This neutral behavior with $p = 0$ is observed because the simple model by Eq. (B1) neglects the effect of product inhibition, which causes the convergence of $\alpha$ and $\beta$ to the same value.

## APPENDIX C: SPATIAL PATTERN

The pattern formation observed in the simulation [Fig. 7(b)] can be modeled by adding a diffusional term to Eq. (B1):

$$\frac{\partial}{\partial t}\begin{pmatrix}\alpha(x,t)\\ \beta(x,t)\end{pmatrix} = k'(S-\alpha-\beta)^2\begin{pmatrix}(1+pL_\alpha)\alpha\\ (1+pL_\beta)\beta\end{pmatrix} - d\begin{pmatrix}\alpha\\ \beta\end{pmatrix} + D\frac{\partial^2}{\partial x^2}\begin{pmatrix}\alpha\\ \beta\end{pmatrix}. \quad \text{(C1)}$$

The linear-stability analysis shows that the uniform state is unstable when $p > 0$ and a spatial pattern forms spontaneously if the wave number is less than $\sqrt{kp\alpha_0(S-2\alpha_0)/D}$ (Supplemental Material, Sec. S7.2 [30]). This inequality suggests that the pattern forms only when $p > 0$, which is consistent with our observation in the simulation [Fig. 7(b)]. The pattern becomes finer with smaller diffusion and a larger reaction rate. The spatial instability is caused by the inherent instability of the reaction in the sequence space and is not a spatial effect. The formed spatial patterns are stabilized by the frequency-dependent selection, which suppresses the growth of other sequence patterns diffusing from a neighboring pattern.


[1] C. Tuerk and L. Gold, *Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase*, Science **249**, 505 (1990).

[2] A. D. Ellington and J. W. Szostak, *In Vitro Selection of RNA Molecules that Bind Specific Ligands*, Nature **346**, 818 (1990).

[3] G. F. Joyce, *The Antiquity of RNA-Based Evolution*, Nature (London) **418**, 214 (2002).

[4] W. Gilbert, *Origin of Life: The RNA World*, Nature (London) **319**, 618 (1986).

[5] D. P. Bartel and J. W. Szostak, *Isolation of New Ribozymes from a Large Pool of Random Sequences*, Science **261**, 1411 (1993).

[6] J. Attwater, A. Wochner, and P. Holliger, *In-Ice Evolution of RNA Polymerase Ribozyme Activity*, Nat. Chem. **5**, 1011 (2013).

[7] K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, and T. R. Cech, *Self-Splicing RNA: Autoexcision and Autocyclization of the Ribosomal RNA Intervening Sequence of Tetrahymena*, Cell **31**, 147 (1982).

[8] P. G. Higgs and N. Lehman, *The RNA World: Molecular Cooperation at the Origins of Life*, Nat. Rev. Genet. **16**, 7 (2015).

[9] D. P. Horning and G. F. Joyce, *Amplification of RNA by an RNA Polymerase Ribozyme*, Proc. Natl. Acad. Sci. U.S.A. **113**, 9786 (2016).

[10] H. Mutschler, A. Wochner, and P. Holliger, *Freeze–Thaw Cycles as Drivers of Complex Ribozyme Assembly*, Nat. Chem. **7**, 502 (2015).

[11] M. W. Powner, B. Gerland, and J. D. Sutherland, *Synthesis of Activated Pyrimidine Ribonucleotides in Prebiotically Plausible Conditions*, Nature (London) **459**, 239 (2009).

[12] F. R. Bowler, C. K. Chan, C. D. Duffy, B. Gerland, S. Islam, M. W. Powner, J. D. Sutherland, and J. Xu, *Prebiotically Plausible Oligoribonucleotide Ligation Facilitated by Chemoselective Acetylation*, Nat. Chem. **5**, 383 (2013).

[13] T. Walton and J. W. Szostak, *A Highly Reactive Imidazolium-Bridged Dinucleotide Intermediate in Nonenzymatic RNA Primer Extension*, J. Am. Chem. Soc. **138**, 11996 (2016).

[14] C. Deck, M. Jauker, and C. Richert, *Efficient Enzyme-Free Copying of All Four Nucleobases Templated by Immobilized RNA*, Nat. Chem. **3**, 603 (2011).

[15] T. A. Lincoln and G. F. Joyce, *Self-Sustained Replication of an RNA Enzyme*, Science **323**, 1229 (2009).

[16] N. Vaidya, M. L. Manapat, I. A. Chen, R. Xulvi-Brunet, E. J. Hayden, and N. Lehman, *Spontaneous Network Formation among Cooperative RNA Replicators*, Nature (London) **491**, 72 (2012).

[17] H. Griesser, P. Tremmel, E. Kervio, C. Pfeffer, U. E. Steiner, and C. Richert, *Ribonucleotides and RNA Promote Peptide Chain Growth*, Angew. Chem., Int. Ed. **56**, 1219 (2017).

[18] J. Derr, M. L. Manapat, S. Rajamani, K. Leu, R. Xulvi-Brunet, I. Joseph, M. A. Nowak, and I. A. Chen, *Prebiotically Plausible Mechanisms Increase Compositional Diversity of Nucleic Acid Sequences*, Nucl. Acids Res. **40**, 4711 (2012).

[19] T. Walton and J. W. Szostak, *A Kinetic Model of Nonenzymatic RNA Polymerization by Cytidine-5′-Phosphoro-2-Aminoimidazolide*, Biochem. **56**, 5739 (2017).

[20] C. B. Mast, S. Schink, U. Gerland, and D. Braun, *Escalation of Polymerization in a Thermal Gradient*, Proc. Natl. Acad. Sci. U.S.A. **110**, 8030 (2013).

[21] H. F. Blum, *On the Origin and Evolution of Living Machines*, Am. Sci. **49**, 474 (1961).

[22] M. Eigen and P. Schuster, *The Hypercycle*, Naturwissenschaften **65**, 7 (1978).

[23] D. R. Mills, R. L. Peterson, and S. Spiegelman, *An Extracellular Darwinian Experiment with a Self-Duplicating Nucleic Acid Molecule*, Proc. Natl. Acad. Sci. U.S.A. **58**, 217 (1967).

[24] C. O. Wilke, *Quasispecies Theory in the Context of Population Genetics*, BMC Evol. Biol. **5**, 44 (2005).

[25] C. O. Wilke, J. L. Wang, C. Ofria, R. E. Lenski, and C. Adami, *Evolution of Digital Organisms at High Mutation Rates Leads to Survival of the Flattest*, Nature (London) **412**, 331 (2001).

[26] M. Kreysing, L. Keil, S. Lanzmich, and D. Braun, *Heat Flux across an Open Pore Enables the Continuous Replication and Selection of Oligonucleotides towards Increasing Length*, Nat. Chem. **7**, 203 (2015).

[27] E. Szathmary, *Simple Growth Laws and Selection Consequences*, TREE **6**, 366 (1991).

[28] R. Ehricht, T. Ellinger, and J. S. Mcaskill, *Cooperative Amplification of Templates by Cross-Hybridization (CATCH)*, FEBS J. **243**, 358 (1997).

[29] T. Kirner, J. Ackermann, D. Steen, R. Ehricht, T. Ellinger, P. Foerster, and J. S. McCaskill, *Complex Patterns in a Trans-Cooperatively Coupled DNA Amplification System*, Chem. Eng. Sci. **55**, 245 (2000).

[30] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.9.011056 for additional details on both the experiments and the theoretical modeling. It is structured as follows: Sec. S1. Ligation reactions. Experimental protocols, sequences, and melting temperature measurements; Sec. S2. Quantitative PCR.

Protocol and calibration of 40 mer two-letter quantification with cold PCR; Sec. S3. Dissociation constant and ligation rate. Kinetic experiments to determine the ligation rate; Sec. S4. Gel electrophoresis. Determination of the length distribution with denaturing electrophoresis; Sec. S5. Numerical simulations. Strategy, details, and examples of the numerical calculation; Sec. S6. Stochastic emergence of self-sustaining sequence structure. Documentation of numerical calculations demonstrating the symmetry breaking in bulk solutions; Sec. S7. Theoretical analysis. One-dimensional analytical models of competitive higher-order growth; and Sec. S8. Supplementary experimental data. Documentation of replicates of the experimental results.

[31] C. Gibard, S. Bhowmik, M. Karki, E. K. Kim, and R. Krishnamurthy, *Phosphorylation, Oligomerization and Self-Assembly in Water under Potential Prebiotic Conditions*, Nat. Chem. **10**, 212 (2018).

[32] F. Barany, *Genetic Disease Detection and DNA Amplification Using Cloned Thermostable Ligase*, Proc. Natl. Acad. Sci. U.S.A. **88**, 189 (1991).

[33] C. B. Mast and D. Braun, *Thermal Trap for DNA Replication*, Phys. Rev. Lett. **104**, 188102 (2010).

[34] A. V. Tkachenko and S. Maslov, *Spontaneous Emergence of Autocatalytic Information-Coding Polymers*, J. Chem. Phys. **143**, 045102 (2015).

[35] C. Fernando, G. Von Kiedrowski, and E. Szathmary, *A Stochastic Model of Nonenzymatic Nucleic Acid Replication: "Elongators" Sequester Replicators*, J. Mol. Evol. **64**, 572 (2007).

[36] J. S. McCaskill, *Spatially Resolved in vitro Molecular Ecology*, Biophys. Chem. **66**, 145 (1997).

[37] P. Baaske, F. M. Winert, S. Duhr, K. H. Lemke, M. J. Russel, and D. Braun, *Extreme Accumulation of Nucleotides in Simulated Hydrothermal Pore Systems*, Proc. Natl. Acad. Sci. U.S.A. **104**, 9346 (2007).

[38] E. Szathmáry, *On the Propagation of a Conceptual Error Concerning Hypercycles and Cooperation*, J. Sys. Chem. **4**, 1 (2013).

[39] K. Harada and L. E. Orgel, *Unexpected Substrate Specificity of T4 DNA Ligase Revealed by in vitro Selection*, Nucl. Acids Res. **21**, 2287 (1993).

[40] K. Harada and L. E. Orgel, *In vitro Selection of Optimal DNA Substrates for Ligation by a Water-Soluble Carbodiimide*, J. Mol. Evol. **38**, 558 (1994).

[41] A. S. Tupper and P. G. Higgs, *Error Thresholds for RNA Replication in the Presence of Both Point Mutations and Premature Termination Errors*, J. Theor. Biol. **428**, 34 (2017).

[42] J. A. Shay, C. Huynh, and P. G. Higgs, *The Origin and Spread of a Cooperative Replicase in a Prebiotic Chemical System*, J. Theor. Biol. **364**, 249 (2015).

[43] P. W. Anderson, *Suggested Model for Prebiotic Evolution: The Use of Chaos*, Proc. Natl. Acad. Sci. U.S.A. **80**, 3386 (1983).

[44] P. R. Wills, S. A. Kauffman, B. M. R. Stadler, and P. F. Stadler, *Selection Dynamics in Autocatalytic Systems: Templates Replicating through Binary Ligation*, Bull. Math. Biol. **60**, 1073 (1998).

[45] S. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford University, New York, 1993).

[46] R. J. Bagley and J. D. Farmer, in *Spontaneous Emergence of a Metabolism*, Alife II edited by C. G. Langton, C. Tayler, J. D. Farmer, and S. Rasmussen (Addison-Wesley, New York, 1991), pp. 93–140.

[47] R. J. Bagley, J. D. Farmer, and W. Fontana, in *Evolution of a Metabolism*, Alife II edited by C. G. Langton, C. Tayler, J. D. Farmer, and S. Rasmussen (Addison-Wesley, New York, 1991), pp. 141–158.

[48] E. G. Worst, P. Zimmer, E. Wollrab, K. Kruse, and A. Ott, *Unbounded Growth Patterns of Reproducing, Competing Polymers—Similarities to Biological Evolution*, New. J. Phys. **18**, 103003 (2016).

[49] C. Bresch, U. Niesert, and D. Harnasch, *Hypercycles, Parasites, and Packages*, J. Theor. Biol. **85**, 399 (1980).

[50] M. C. Boerlijst and P. Hogeweg, *Spiral Waves Structure in Pre-Biotic Evolution: Hypercycles Stable against Parasites*, Physica (Amsterdam) **48D**, 17 (1991).

[51] A. S. Zadorin, Y. Rondelez, J. C. Galas, and A. Estevez-Torres, *Synthesis of Programmable Reaction-Diffusion Fronts Using DNA Catalyzers*, Phys. Rev. Lett. **114**, 068301 (2015).

[52] C. K. Biebricher and R. Luce, *In vitro Recombination and Terminal Elongation of RNA by Q Beta Replicase*, EMBO J. **11**, 5129 (1992).

[53] J. S. McCaskill and G. J. Bauer, *Images of Evolution: Origin of Spontaneous RNA Replication Waves*, Proc. Natl. Acad. Sci. U.S.A. **90**, 4191 (1993).

[54] D. Braun and A. Libchaber, *Trapping of DNA by Thermophoretic Depletion and Convection*, Phys. Rev. Lett. **89**, 188103 (2002).

[55] L. Keil, M. Hartmann, S. Lanzmich, and D. Braun, *Probing of Molecular Replication and Accumulation in Shallow Heat Gradients through Numerical Simulations*, Phys. Chem. Chem. Phys. **18**, 20153 (2016).

[56] K. Soai, T. Shibata, H. Morioka, and K. Choji, *Asymmetric Autocatalysis and Amplification of Enantiomeric Excess of a Chiral Molecule*, Nature (London) **378**, 767 (1995).

[57] J. V. Gavette, M. Stoop, N. V. Hud, and R. Krishnamurthy, *RNA–DNA Chimeras in the Context of an RNA World Transition to an RNA/DNA World*, Angew. Chem., Int. Ed. **55**, 13204 (2016).

[58] A. Mariani and J. D. Sutherland, *Non-Enzymatic RNA Backbone Proofreading through Energy-Dissipative Recycling*, Angew. Chem. **129**, 6663 (2017).

[59] M. Kramer and C. Richert, *Enzyme-Free Ligation of 5′-Phosphorylated Oligodeoxynucleotides in a DNA Nanostructure*, Chem. Biodiv. **14**, e1700315 (2017).

[60] M. Kalinowski, R. Haug, H. Said, S. Piasecka, M. Kramer, and C. Richert, *Phosphoramidate Ligation of Oligonucleotides in Nanoscale Structures*, Chembiochem **17**, 1150 (2016).

[61] V. Patzke, J. S. McCaskill, and G. von Kiedrowski, *DNA with 3′–5′-Disulfide Links—Rapid Chemical Ligation through Isosteric Replacement*, Angew. Chem., Int. Ed. **53**, 4222 (2014).

[62] C. He, I. Gállego, B. Laughlin, M. A. Grover, and N. V. Hud, *A Viscous Solvent Enables Information Transfer from Gene-Length Nucleic Acids in a Model Prebiotic Replication Cycle*, Nat. Chem. **9**, 318 (2017).

[63] H. Mutschler, A. Wochner, and P. Holliger, *Freeze–Thaw Cycles as Drivers of Complex Ribozyme Assembly*, Nat. Chem. **7**, 502 (2015).

[64] A. E. Engelhart, K. P. Adamala, and J. W. Szostak, *A Simple Physical Mechanism Enables Homeostasis in Primitive Cells*, Nat. Chem. **8**, 448 (2016).

[65] J. A. Yeates, C. Hilbe, M. Zwick, M. A. Nowak, and N. Lehman, *Dynamics of Prebiotic RNA Reproduction Illuminated by Chemical Game Theory*, Proc. Natl. Acad. Sci. U.S.A. **113**, 5030 (2016).