Kinetic Selection of Template Polymer with Complex Sequences

Yoshiya J. Matsubara[*] and Kunihiko Kaneko[†]

Graduate School of Arts and Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan

The emergence and maintenance of polymers with complex sequences pose a major question in the study of the origin of life. To answer this, we study a model polymerization reaction, where polymers are synthesized by stepwise ligation from two types of monomers, catalyzed by a long polymer as a template. Direct stochastic simulation and dynamical systems analysis reveal that the most dominant polymer sequence in a population successively changes, depending on the flow rate of monomers to the system, with more complex sequences selected at a lower flow rate. We discuss the relevance of this kinetic sequence selection through nonequilibrium flow to the origin of complex polymers.

Schrödinger, in his celebrated monograph "What is Life" [1], recognized that the essence of a heredity carrier lies in aperiodic crystals, as was soon confirmed by the elucidation of DNA structure. In general, biopolymers consist of different kinds of monomers, constituting an aperiodic complex sequence, as is important not only for genetic information but also for the catalytic function.

With regard to the origin of life or self-replication, catalytic polymers are required, as has also been investigated in recent experiments [2–6]. In their study, polymers with complex sequences had to be synthesized and preserved, to encode a large amount of information and for catalytic functions [7,8]. Hence, it is important to uncover the condition that allows for the generation of a polymer with a complex monomer sequence.

Polymers such as RNA are replicated using other polymers as templates. In the prebiotic world, complex monomer sequences therefore have been synthesized by reactions with templates as catalysts, which are also synthesized via such template reaction. Thus, the polymerization processes are autocatalytic in nature, as has been extensively investigated [9–16], including theoretical models that explicitly consider templatelike self-replicating polymerization [8,17–24] and some synthesis experiments [6,25,26].

If there exist multiple catalytic polymers with sufficient lengths, then the autocatalytic process for their synthesis can select one of such polymers. Here we study the dependence of the selected sequence on the flow rate of monomers. The replication of polymers needs a supply of monomers to compensate polymer degradation. Such a

nonequilibrium open condition could be provided by a hydrothermal vent at the origin of life and by a chemostat condition in a laboratory experiment [27]. Despite the recognition that such nonequilibrium flow is essential to the emergence and maintenance of life, its influence on the selection of sequence has not been systematically investigated.

The selection of a specific polymer by catalytic reactions studied so far [8,17–24] does not show any dependence on external conditions. Reference [16] reported the dependence of the selected polymer sequence on the bias in the monomer components. In contrast, here, we study a template-catalyst polymerization model by a complementary sequence and demonstrate that the selected sequence of the template polymer depends critically on the flow rate, and more complex sequences are selected as the flow rate is decreased. This selection mechanism is the *kinetic*, rather than *energetic*, selection of complex sequences (that include a variety of subsequences, which are defined later) without any specific design of energy dependence, as a result of changing the rate of external supply of monomers. As will be analyzed by dynamical systems and combinatorial analysis, complex sequences that are synthesized via diverse kinetic pathways from monomers are selected for a low flow rate. The generality of this selection mechanism will be discussed with possible relevance to prebiotic polymer synthesis.
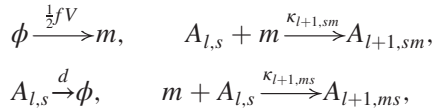
We consider a synthetic reaction of polymers, which consist of two kinds of monomers, denoted as 0 and 1, where each polymerization is catalyzed by a long template polymer. Thus, all polymer species are described by the binary integer s of length l (e.g., 01, 0101, 00000, etc.). The polymerization progresses in a container with volume V under a well-mixed (homogeneous) condition. Only the two monomer species flow into the container from outside at the same constant rate $\frac{1}{2}fV$, while all the molecules in the container diffuse out at rate d.

118101-1

For simplicity, only the ligation reaction between the polymers and monomers is considered here. The polymer has directionality, and the monomer can ligate from both the left and right sides of a polymer (e.g., $1 + 00 \rightarrow 100$ and $00 + 1 \rightarrow 001$). For simplicity and to focus on the kinetic aspect, all the polymer species of the same length are assumed to have the same chemical potential. Thus, in equilibrium, the concentrations of all the polymer species with the same length are equal.

A template polymer serving as a catalyst accelerates both the forward and backward reactions, without changing the equilibrium condition. The catalytic reaction works if the template includes a subsequence of the bit inversion of the product. For example, the reaction $111 + 1 \rightarrow 1111$ is catalyzed by the template containing 0000 as a subsequence, i.e., 00000, 10000, or 00001. For simplicity, only the longest polymers with the given length L, independent of their sequence, can serve as the template in ligation reactions. Since all the ligation reactions do not change the total number of monomers, the ratio of monomer inflow to the outflow by the diffusion of the polymers determines the total number of monomers in the system; i.e., the total monomer concentration is given by f/d.

In summary, the chemical reactions are given as follows by denoting the polymer of the length l with the sequence $s = \{0,1\}^l$ as $A_{l,s}$:

$$\phi \xrightarrow{\frac{1}{2}fV} m, \qquad A_{l,s} + m \xrightarrow{\kappa_{l+1,sm}} A_{l+1,sm},$$

$$A_{l,s} \xrightarrow{d} \phi, \qquad m + A_{l,s} \xrightarrow{\kappa_{l+1,ms}} A_{l+1,ms},$$

where m is a monomer 0 or 1 and sm and ms are the sequences with m added to the left side or right side of s, respectively [28,29]. The polymerization can occur spontaneously at a rate ϵ, which is set to be quite small.

Under the large volume limit, the concentration of polymer $A_{l,s}$, $x_{l,s}$ follows the deterministic rate equation:

$$\dot{x}_{1,m} = \frac{1}{2}f - x_{1,m}\sum_{l',s'}(\kappa_{l'+1,s'm} + \kappa_{l'+1,ms'})x_{l',s'} - dx_{1,m},$$

$$\dot{x}_{l,s} = \kappa_{l,s}x_{1,m}(x^{(1)}_{l-1,s} + x^{(2)}_{l-1,s})$$
$$- x_{l,s}\sum_{m'}x_{1,m'}(\kappa_{l+1,sm'} + \kappa_{l+1,m's}) - dx_{l,s},$$

$$\dot{x}_{L,s} = x_{1,m}x_{L,s}(x^{(1)}_{L-1,s} + x^{(2)}_{L-1,s}) - dx_{L,s}. \qquad (1)$$

Here, $x^{(1)}_{l-1,s}$ and $x^{(2)}_{l-1,s}$ are the concentrations of the $(l-1)$-mer, which is a subsequence of $A_{l,s}$ on both sides [30], $\kappa_{l,s} = \frac{1}{2}(\epsilon + \sum_{(l',s')\in\mathcal{T}_{l+1,s}} n_{l',s'}/V)$, and $\mathcal{T}_{l,s}$ is the set of all template molecules that can catalyze the reactions to produce the polymer species $A_{l,s}$. This equation has multiple attractors: As the pair of the longest polymer of a given sequence and its complementary pair catalyze the
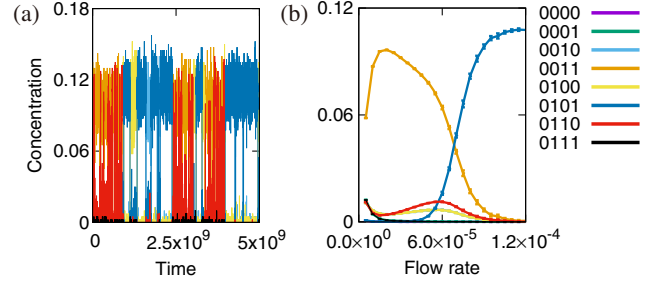

FIG. 1. (a) Time series of concentrations of all template species. $L = 4$, $V = 400$, $f/d = 1$, $\epsilon = 10^{-4}$, and $f = 7.0 \times 10^{-5}$. (b) Average concentration of each template sequence in the stationary state. The horizontal axis is the flow rate in chemostat $f(=d)$. The time intervals for average are varied from 10^{13} to 10^{14} depending on f. A sequence and its left-right reverse take the same average concentration over a long time, and the corresponding two lines are overlapped.

reaction to synthesize them, the concentrated population for each of the complementary pairs can be an attractor. It should be noted that, since the complementary sequence works as a template for the polymerization, the complementary pair (e.g., 0010 and 1101) always coexists in an equal fraction.

If the system size is finite, fluctuations around the rate equation (1) induce switching among the attractors, and specific attractor(s) are selected dominantly. We investigated this selection by numerically solving the original stochastic reaction model [31]. In Fig. 1(a), the time series of the concentration of template species is plotted. A specific complementary pair of sequences is dominant for some time span, followed by switching to a state with a different dominant pair. As the concentrations of complementary sequences are equal, only one of them is plotted throughout the Letter.

Next, the temporal average of the concentration of each template sequence was computed to examine its dependence on the flow rate f [Fig. 1(b)] by fixing $f/d = 1$ (hence, the total monomer concentration f/d is kept constant). As shown, the concentration of each sequence and, in particular, that of the dominant species depend on f. Note that a sequence and its reverse form [e.g., 0010(1101) and 0100(1011)] have the same fraction over a long time average due to symmetry, although, at each instance, one of them is selected in the system. For $L = 4$, 0101 (and 1010) is selected for a large f, while 0011 is selected for a small f. For $L = 6$, as is shown in Supplemental Material [32], Fig. S1(a), the dominant sequence changes as $010101 \rightarrow 011001 \rightarrow 001001 \rightarrow 000100$ with a decrease in f.

These changes in the dominant sequence against the decrease in the flow rate are interpreted as the increase in the sequence complexity. Here, a "complex" sequence implies that it contains more subsequences including their complementary ones. For example, 0101 includes only 01 and 10 dimers, while 0010 includes 00, 01, 10, and 11 and

(a)   (b)
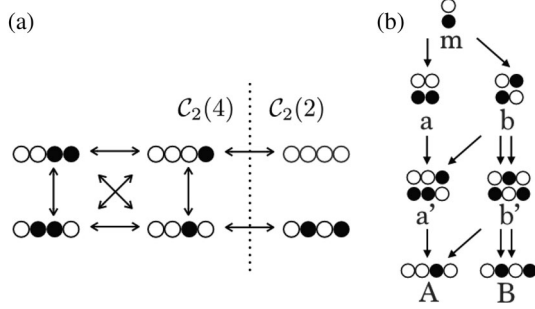

$C_2(4)$: $C_2(2)$

FIG. 2.   (a) Transition network between attractors. A double head indicates a transition between them, and each sequence represents the one of a complementary pair that is dominant at each attractor. (b) Schematic diagram representing the reaction pathway from monomers to each template A and B. White and black circles represent 0 and 1 monomers, respectively. Arrows indicate the ligation reaction of a shorter polymer with a monomer. Each reaction proceeds using A or B or both as a template. Detailed coefficients are in the main text.

is thus more complex. In general, let $C_l(2k)$ be the set of template sequences that have $2k$ species of l-mers as subsequences (note that the complementary sequence is included in the counting). For example, 0101 belongs to $C_2(2)$, because it includes two dimer subsequences 01 and 10, and 0110 belongs to $C_2(4)$, as it includes 01, 11, 10, and 00. To study the selection of a complex sequence with f, we computed the sum of the average residence time at each attractor corresponding to the sequences belonging to $C_l(2k)$. It is defined by $x_l(2k) = \sum_{s \in C_l(2k)} x_{L,s}$. In Fig. S1(b) in Supplemental Material [32], $x_l(2k)$ is plotted as a function of f. For $L = 4$, the dominant concentrations change from $x_2(4)$ to $x_2(2)$, and for $L = 6$, they change in the order $x_3(8)$, $x_3(6)$, $x_3(4)$, $x_3(2)$ with an increase in f, implying that a more complex sequence is selected for a smaller f.

To understand this change in the dominant sequence, we note that the residence time for each attractor, under noise due to the finiteness in the molecule number, is larger if the attraction toward it is stronger. To understand this change in the dominant sequence, we first study the change from the sequence with $C_2(4)$ to that with $C_2(2)$. Since the polymerization progresses from a dimer to a trimer, and then to a tetramer, the transition first occurs between tetramers that share the trimers, i.e., those with only a one-bit difference. The transition diagram is shown in Fig. 2(a), where the change from $C_2(2)$ with 0101 to $C_2(4)$ with a decrease in f is mediated by the transition to sequence 1101 (0010). Hence, we first discuss the competition between the attractor with the dominance of the 0101 and 1010 pair (denoted as B) and that with the 0010 and 1101 pair (as A) by focusing only on these two pairs of templates, while neglecting other template polymers [36]. By assuming that the number of shorter molecules changes faster than do the templates, the reduced rate equation is written only in terms of the concentrations of these templates, as follows:
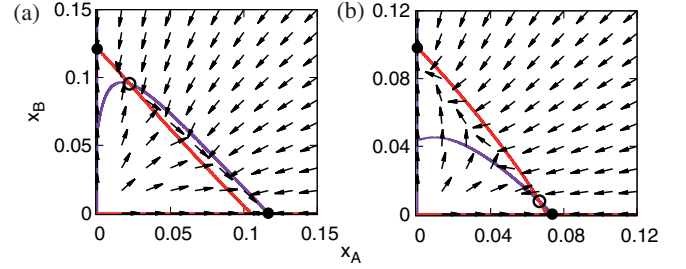
(a)   (b)

FIG. 3.   Vector field of x_A and x_B. The blue line is the nullcline of $\dot{x}_A = 0$, and the red is the nullcline of $\dot{x}_B = 0$. The nullclines overlap the x_A line and x_B line partially. In both figures, there are two stable fixed points (filled circle) and one unstable fixed point (empty circle). Two stable fixed points are A-abundant and B-abundant attractors, respectively. As the value of f increases from $f(=d) = 1.0 \times 10^{-5}$ (a) to 5.0×10^{-4} (b), the unstable fixed point moves toward the A-abundant attractor.

$$\dot{x}_A = r_A(x_A, x_B)x_A - dx_A, \qquad (2)$$

$$\dot{x}_B = r_B(x_A, x_B)x_B - dx_B. \qquad (3)$$

Here, r_A (r_B) is the synthesis rate of the corresponding template for each type A (B) for a given x_A or x_B, respectively, whose form is obtained from Eq. (S3), as shown in Supplemental Material [32]. This rate equation has two fixed-point attractors, $(x_A = x_A^*, x_B = 0)$ and $(x_A = 0, x_B = x_B^*)$ [37]. The vector field for dx_A/dt, dx_B/dt is thus obtained as in Fig. 3. Here, the attractor close to the unstable fixed point is kicked out by noise.

It is shown straightforwardly that the attraction speed to attractor A is given by $v_A \equiv r_B(x_A^*, 0) - r_A(x_A^*, 0)$, from Eqs. (2) and (3) and the eigenvalues of the Jacobian matrix of the dynamics, and that to B is $v_B \equiv r_B(0, x_B^*) - r_A(0, x_B^*)$ (see Supplemental Material [32]). As shown therein, $v_B - v_A < 0$ for $f \sim 0$ and > 0 for $f \gg 1$, explaining the transition from B to A by noise with the decrease in f. This flow rate dependence is also interpreted by the diversity in reaction pathways to replicate templates A and B. While only one reaction pathway exists for synthesizing B from the monomers with double the speed because of symmetry, various pathways exist for the synthesis of A from the monomers, because they have more types of subsequences. For large $f(=d)$, double autocatalytic paths to B work efficiently, while the existence of diverse subsequences for the pathways to A causes a larger loss of shorter polymers, leading to $r_B > r_A$. For small f, the concentrated use of the same dimer for B results in its deficiency, and r_A is larger, since A can use both dimers. (See Supplemental Material [32] for an analytic derivation of this dependence).

The above argument holds for all the template sequences belonging to $C_2(2)$ and $C_2(4)$. Therefore, the total concentration of the sequence $x_2(2)$ [$x_2(4)$] is large when f is large (small). Furthermore, it is valid for the L-mer template
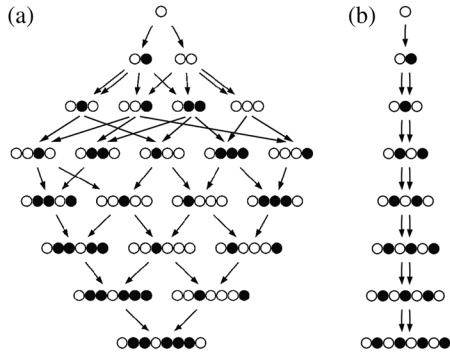
FIG. 4. Reaction pathway to synthesize 8-mer templates from monomers for (a) a complex sequence [01101110 $\in \mathcal{C}_4(10)$] and (b) a simple sequence [01010101 $\in \mathcal{C}_4(2)$]. Black and white circles represent monomers 1 and 0, respectively. Each arrow indicates a ligation reaction between a shorter polymer and a monomer, which are catalyzed by product templates.

sequence A belonging to $\mathcal{C}_l(2k+2)$ and B belonging to $\mathcal{C}_l(2k)$ sharing subsequences in part. Here again, the replication speeds of A and B, i.e., r_A and r_B, determine the attraction speed to the fixed points A and B, and the attractor A with a complex sequence is dominant for a low f. In Fig. 4, the ligation reactions from the monomer to 8-mer templates are drawn. The most complex sequence with $\mathcal{C}_4(10)$ has a variety of pathways, as compared with the simple sequence with $\mathcal{C}_4(2)$. For a low f, the sequence with $\mathcal{C}_4(10)$ is expected to be selected, while for a high f, that with $\mathcal{C}_4(2)$ is selected. The fraction of each sequence obtained from a direct simulation is plotted in Fig. 5 as a function of f.

Since the direct simulation of cases with larger L is time consuming, we studied a reduced model, where only the
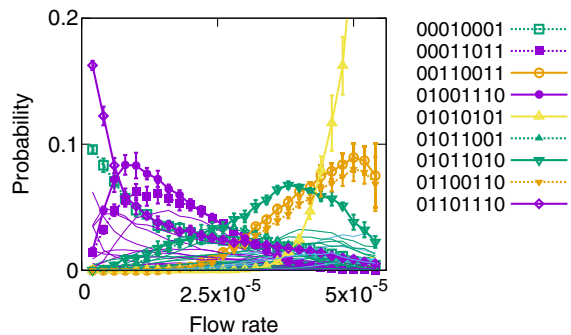

FIG. 5. Residence probability of an attractor at which a pair of sequences dominates. We set the length of the template L as 8 and $\epsilon = 5.0 \times 10^{-6}$, $V = 500$. The major sequences are plotted as thick lines. The probability is calculated as the time average of 10^{14}–10^{15} simulation times. One of a reverse pair (e.g., 01101110 and 01110110) is omitted here. The complexity of the dominant sequence decreases with a decreasing flow rate: The dominant sequence changes from 01010101 [$\mathcal{C}_4(2)$], 00110011 [$\mathcal{C}_4(4)$], and 01011010 [$\mathcal{C}_4(6)$] to 01001110 or 01101110 [$\mathcal{C}_4(10)$] with a decrease in the flow rate.

transition between the attractors with a certain dominant template was considered, with the transition probability determined by the difference between the replication rates of the two templates (see Supplemental Material [32] for details). A complex sequence with a large number of subsequences was selected as the flow rate decreased.

We have confirmed the generality of the present result, i.e., a successive increase in the complexity of the dominant sequence with the decrease in the flow rate, for the following cases (see Supplemental Material [32]): (i) inclusion of variation in the catalytic activity by each template, if it is not too large; (ii) change in the rules for catalytic strength; (iii) inclusion of the ligation between polymers longer than monomers; (iv) difference in the concentrations of monomers (0,1)—in the study here, complementary polymers take the same concentrations, irrespective of the monomer concentrations, and, thus, the bias in monomer cannot select one of the complementary molecules; (v) increase in spontaneous ligation rate ϵ. Finally, in the study presented here, the number of monomer types is only 2, in contrast to 4 in RNAs and 20 amino acids in proteins. Still, the increase in polymer complexity by the diversity of reaction pathways holds for a larger number of monomer types (see also [38]).

In summary, we have shown that, in a polymerization process with template-based catalytic reactions, the dominant sequence selected depends on the flow rate of the monomers, and its complexity increases with the decrease in the rate. The selection is based on two basic mechanisms: (I) the attractor selection of a higher growth rate due to stochasticity in reactions and (II) the preference of a complex sequence to have a variety of pathways to avoid the deficiency of subpart shorter polymers.

(I) The finiteness of the system size provides stochasticity (noise) in a reaction, which gives rise to switches among attractors with different dominant sequences. Attractors with a higher growth rate have a larger stability against noise (see, e.g., Fig. 3) and are selected dominantly under the presence of noise. If the noise strength is too small, no transitions occur among the attractors, whereas if it is too large, no specific sequence dominates [39] (see also [40,41] for attractor selection of a higher cellular growth and [14,42,43] for the relevance of noise to chemical evolution).

(II) When the flow rate is limited, the polymerization of a simple (say, 01-periodic) sequence suffers from the deficiency of subsequence polymers of shorter length. The polymerization of a complex sequence of diverse pathways can keep a variety of subsequence polymers (see also [44] for an increase in component diversity in the protocell under low nutrient flow), while under a larger d value, the decomposition of many components is disadvantageous for the growth.

This variety of subsequence leads to the definition of the complexity in sequence we adopted here [45]. Of course,

this is one possible definition, whereas a related definition is adopted to compare the subsequences in the genome in bioinformatics [46–51]. The relationship between the sequence complexity and structural or functional property of polymers has been discussed in several studies [52]. In our study, such a complex sequence is kinetically selected in the low flow rate region.

The results presented here will provide a novel perspective on kinetic conditions for the origin of life. Although a nonequilibrium condition is needed for life, our result implies that an excessively strong nonequilibrium condition will damage the sequence complexity. For the emergence and maintenance of prebiotic autocatalytic systems, optimal nonequilibrium conditions are required. Although it is difficult to estimate the actual flow rate for the appearance of complex sequences, the flux rate dependence of the complexity of synthesized polymers will be important to uncover the possible condition for the origin of life and for the laboratory construction of prebiotic self-replication systems.

---

[*] yoshi@copmlex.c.u-tokyo.ac.jp

[†] kaneko@complex.c.u-tokyo.ac.jp

[1] E. Schrödinger, *What is Life?* (Cambridge University Press, Cambridge, England, 1943).

[2] G. von Kiedrowski, Angew. Chem., Int. Ed. Engl. **25**, 932 (1986).

[3] T. A. Lincoln and G. F. Joyce, Science **323**, 1229 (2009).

[4] H. Krammer, F. M. Möller, and D. Braun, Phys. Rev. Lett. **108**, 238104 (2012).

[5] N. Vaidya, M. L. Manapat, I. A. Chen, R. Xulvi-Brunet, E. J. Hayden, and N. Lehman, Nature (London) **491**, 72 (2012).

[6] J. W. Sadownik, E. Mattia, P. Nowak, and S. Otto, Nat. Chem. **8**, 264 (2016).

[7] J. M. Carothers, S. C. Oestreich, J. H. Davis, and J. W. Szostak, J. Am. Chem. Soc. **126**, 5130 (2004).

[8] J. Derr, M. L. Manapat, S. Rajamani, K. Leu, R. Xulvi-Brunet, I. Joseph, M. A. Nowak, and I. A. Chen, Nucleic Acids Res. **40**, 4711 (2012).

[9] M. Eigen and P. Schuster, Naturwissenschaften **64**, 541 (1977).

[10] J. D. Farmer, S. A. Kauffman, and N. H. Packard, Physica D (Amsterdam) **22**, 50 (1986).

[11] P. Szabó, I. Scheuring, T. Czárán, and E. Szathmáry, Nature (London) **420**, 340 (2002).

[12] K. Kaneko, Adv. Chem. Phys. **130**, 543 (2005).

[13] V. Giri and S. Jain, PLoS One **7**, e29546 (2012).

[14] Y. J. Matsubara and K. Kaneko, Phys. Rev. E **93**, 032503 (2016).

[15] E. Guseva, R. N. Zuckermann, and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **114**, E7460 (2017).

[16] G. Kinsler, S. Sinai, N. K. Lee, and M. A. Nowak, PLoS One **12**, e0180208 (2017).

[17] P. W. Anderson, Proc. Natl. Acad. Sci. U.S.A. **80**, 3386 (1983).

[18] A. V. Tkachenko and S. Maslov, J. Chem. Phys. **143**, 045102 (2015).

[19] D. L. Stein and P. W. Anderson, Proc. Natl. Acad. Sci. U.S.A. **81**, 1751 (1984).

[20] C. Fernando, G. Von Kiedrowski, and E. Szathmáry, J. Mol. Evol. **64**, 572 (2007).

[21] M. A. Nowak and H. Ohtsuki, Proc. Natl. Acad. Sci. U.S.A. **105**, 14924 (2008).

[22] H. Ohtsuki and M. Nowak, Proc. R. Soc. B **276**, 3783 (2009).

[23] B. Obermayer, H. Krammer, D. Braun, and U. Gerland, Phys. Rev. Lett. **107**, 018101 (2011).

[24] S. Tanaka, H. Fellermann, and S. Rasmussen, Europhys. Lett. **107**, 28004 (2014).

[25] C. He, I. Gállego, B. Laughlin, M. A. Grover, and N. V. Hud, Nat. Chem. **9**, 318 (2017).

[26] S. Toyabe and D. Braun, arXiv:1802.06544.

[27] A. Novick and L. Szilard, Science **112**, 715 (1950).

[28] For $l = 1$, the top and the bottom reactions are considered only once.

[29] The reverse ligation reaction is neglected here, but its inclusion does not alter the following results qualitatively, as long as its rate is small.

[30] For $l = 2$, Eq. (1) should be read with $x_{1,s}^{(1)} = x_{1,m}$ and $x_{1,s}^{(2)} = 0$.

[31] D. T. Gillespie, J. Phys. Chem. **81**, 2340 (1977).

[32] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.121.118101 for figures with additional information on the case with a hexamer, a detailed analysis in the case of a tetramer, a reduced description of the model, and an examination of the generality of the result, which includes Refs. [33–35].

[33] C. Gardiner, *Stochastic Methods* (Springer-Verlag, Berlin, 1985).

[34] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, New York, 1992), Vol. 1.

[35] M. Eigen, Naturwissenschaften **58**, 465 (1971).

[36] This approximation is valid when the rate of spontaneous ligation is much smaller than the timescale of the transition dynamics.

[37] To analyze the stability of each attractor, the (tiny) spontaneous ligation term is not necessary, and it is neglected.

[38] A. V. Tkachenko and S. Maslov, arXiv:1710.06385.

[39] Spontaneous reaction that undergoes without templates plays a similar role as the noise: If it is too large, there emerges no dominant sequence, and if it is too small, there is no transition between the attractors [32].

[40] A. Kashiwagi, I. Urabe, K. Kaneko, and T. Yomo, PLoS One **1**, e49 (2006).

[41] C. Furusawa and K. Kaneko, PLoS Comput. Biol. **4**, e3 (2008).

[42] F. Jafarpour, T. Biancalani, and N. Goldenfeld, Phys. Rev. Lett. **115**, 158101 (2015).

[43] K. Kaneko and T. Yomo, J. Theor. Biol. **214**, 563 (2002).

[44] A. Kamimura and K. Kaneko, Phys. Rev. E **93**, 062419 (2016).

[45] The definition we adopted is closer to algorithmic complexity. However, it is not concerned with the computer algorithm and is with the diversity of chemical-reaction paths for polymerization.

[46] E. N. Trifonov, Struct. Methods **1**, 69 (1990).

[47] O. Popov, D. Segal, and E. N. Trifonov, BioSystems **38**, 65 (1996).

[48] Y. L. Orlov and V. N. Potapov, Nucleic Acids Res. **32**, W628 (2004).

[49] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).

[50] A. N. Kolmogorov, Sankhyā: Indian J. Stat. Ser. A **25**, 369 (1963).

[51] S. Lloyd and H. Pagels, Ann. Phys. (N.Y.) **188**, 186 (1988).

[52] A. Gabrielian and A. Bolshoy, J. Comput. Chem. **23**, 263 (1999).