博士論文

Emergence and Selection of Self-replicating Polymer System for the Origins of Life

（生命の起源のための自己複製高分子系の出現と選択）

松原嘉哉

Contents

Chapter 1

Introduction

The primary purpose of this thesis is research into the origins of life.

Life is a self-replicating system composed of a wide variety of biopolymers.[1] Major biopolymers include polynucleotides such as RNA, DNA, proteins, lipid polymers. Generally, proteins function as enzymes, DNA records information as a sequence of the bases, and lipids form a double-layer membrane. Proteins are synthesized by ribosomes, large biopolymers composed mainly of RNA, whereas polymerases (enzymes) replicate RNA and DNA. In this way, polymers help each other's synthesis and grow as a whole living system.

The reproduction process in a living system is not perfect, and there are variations. Among life forms, those that can reproduce themselves faster eventually survive. This is the process of evolution that Charles Darwin proposed in his celebrated book "The Origin of Species." [7] The diversity of life on Earth today have been created by adaptively evolving to various environments. Serving as a well-known genetic material in a living system is a long DNA polymer, whose structure was elucidated by the discovery of the double-helix conformation by Watson and Crick in 1953 [8]. Then, the leading cause of error in the replication of life is miscopying of a base in a DNA sequence. The search for the ancestor life form in the current data is carried out by means of the phylogenetic tree of life, which was derived by Carl Woose using differences in the gene that codes for the ribosome polymer as a molecular clock [9]. It was suggested that all species of living creatures have a last universal common ancestor (LUCA). For example, it is predicted that the LUCA has a genetic system that transcribes DNA sequence to mRNA and translates mRNAs to proteins. It is also predicted that the LUCA is a thermophile, and these creatures lived in places like hydrothermal vents in the deep sea. On the other hand, little is known about life before the LUCA because there is almost no trace of these life forms.

[1] There are various alternative definitions of life [1, 2, 3]. If we consider a self-replicating system in general, there are various examples of complex systems and artificial-life research, e.g., Neumann's self-replicating machine [4] and life game in computer simulations, which are not restricted by laws of physics. There are also various examples in science fictions [5, 6]. By contrast, in this thesis, we define life as a system of biopolymers that is restricted by the physical laws of thermodynamics and kinetics.

The greatest mystery in the study on the origins of life is how life systems emerged before the LUCA. The idea that life originated from nonliving materials in the primitive Earth's "primordial soup" was first proposed by Darwin too, and the more concrete hypothesis of the chemical reaction process was suggested by Oparin in the "coacervate hypothesis." [10] It has long been known that organic materials can be synthesized from inorganic materials as first uncovered by the experiment on the synthesis of urea by Wöhler in 1828. Besides, in 1950, Miller showed that several kinds of amino acids could be synthesized from the reducing atmosphere of the primitive Earth [11]. It is now believed that simple organic molecules and amino acids, which are basic units of proteins, could be synthesized naturally on the primitive Earth. Nonetheless, there is a big gap between such simple organic molecules and the system of biopolymers that can reproduce itself like the LUCA.

An important hypothesis in the discussion about the emergence of life before the LUCA is the RNA world [12]. As described above, modern life is composed of multiple biopolymers such as proteins and DNA. On the other hand, the RNA world hypothesis postulates that a self-replicating system composed only of RNA molecules preceded the modern forms of life, and later, DNA has taken over the role of the information storage medium for inheritance, whereas proteins the role of catalysts. RNA can replicate sequence information by forming Watson–Crick pairs with the corresponding monomer species and by using other polymers as templates. The existence of a "ribozyme," which has a catalytic ability like enzymes do, was gradually elucidated recently since the discovery of self-cleaving ribozymes by Thomas Cech [13]. Moreover, it is known that the major part of the ribosome, which translates mRNA and synthesizes proteins, is RNA. These are the reasons why the RNA world is supported by evidence. In fact, there is a chapter on the RNA world in the standard molecular biology textbook "Molecular Biology of the Cell." [14] Many researchers think that the RNA world is a reliable hypothesis about the origins of life. Nevertheless, the RNA hypothesis has many problems too; for example, it is relatively more difficult to synthesize monomers of nucleic acids (nucleotides) than amino acids. Thus, the protein world hypothesis in which the emergence of proteins has preceded for RNA is also believed as an alternative for the RNA world hypothesis. Therefore, the debate about the first biopolymers constituting the prelife system is not completely settled.

Whether in the protein world, RNA world or a autocatalytic system of other biopolymers, how huge and complex polymers had been synthesized from the basic units (monomers) is a big question. In vivo, various huge polymeric catalysts such as ribosomes and polymerases cooperate to synthesize or replicate polymers. For example, the central part of a ribosome are nucleotides with a few thousand in length. Generally, functional polymers in living systems are longer than one hundred, although shorter polymers have a small catalytic ability. The presence of such complicated and huge polymers is required for the synthesis of such polymers themselves. Although it is experimentally tested whether long polymers can be synthesized without such a polymeric catalyst, the synthesis has not reached the length and complexity level that would be required for self-replicating systems.

This problem corresponds to the "chicken or the egg" question regarding the origins of life.

In this thesis, we explore how sufficiently complex polymers that constitute self-replicating systems emerged and persisted using mathematical models. In the research on the origins of life and evolutionary biology, mathematical models are often used because the emergence and evolution of life require too long a period for analysis in a laboratory. In particular, we focus on the problem that several steps of a polymerization process are necessary to synthesize long biopolymers.

In the introduction chapter, first, we will briefly review experiments on noncatalytic synthesis of monomers such as amino acids and polymers such as proteins and RNA as well as the attempts to construct autocatalytic polymer systems in synthetic biology and organic chemistry. Next, theoretical studies related to the origins of life will be reviewed, based on (1) the law of the growth and evolution for replicators, (2) emergence of an autocatalytic system, and (3) evolution of the autocatalytic set toward a complex cell. Finally, we will discuss the outline of this study and its scope for understanding the origins of life.

1.1 Experimental implementation for the origins-of-life hypothesis

Biopolymers such as RNA or proteins are composed of many types of monomers [14]. The monomers of nucleotides consist of sugars, bases, and phosphates. There are C (cytosine), G (guanine), A (adenine), and U (uracil) as bases. For proteins, there are 20 types of monomers (e.g., glycine and alanine) in vivo. These monomers can form a covalent bond with each other (the peptide bond for amino acids and the phosphodiester bond for nucleotides) and polymerize into long sequences.

The monomer sequences of biopolymers are regarded as a genotype that is inherited after replication as information and determines the phenotypes such as a function as a catalyst. Catalysts that are proteins are called enzymes and those that are RNAs are ribozymes. The structure of biopolymers determines their functions such as catalytic activity. Both proteins and RNAs can be folded into three-dimensional structure depending on their sequence of monomers (genotype). The tertiary structure of both RNAs and proteins is predicted numerically.

In this section, we will briefly look at attempts to synthesize such monomers and polymers and their autocatalytic sets in organic chemistry or synthetic biology experiments. (see details in review articles in [15, 16, 17, 18, 19].)

1.1.1 Experiments on the synthesis of biopolymers and monomers

Amino acids were synthesized in the Miller–Urey experiment [11]. 23 types of amino acids [20] are easily synthesized by passing an electric current through a reducing gas composed of water, methane, ammonia, and hydrogen, which was

supposed to constitute the atmosphere of the primitive Earth. These experimental conditions are now considered unrealistic for the primitive Earth, and amino acids are synthesized under various alternative conditions. Regarding nucleotides, the routes of synthesis of the sugars (ribose for RNA) and bases (G, C, A, and U), which constitute nucleotides (monomers), have been researched. For example, it has been shown by Oró that purine bases (guanine and cytosine) can be synthesized in the polymerization reaction of HCN [21]. Additionally, the ribose sugar is synthesized in the formose reaction although its yield is quite low [22]. Compared with an amino acid, nucleotides are not so easy to synthesize because the process of assembly of these sugars and bases generates various types of byproducts, and the yield of canonical nucleotides is low. Recently, the alternative pathway to synthesize nucleotides is searched by Sutherland et al., which is not through bases and sugars as intermediates [23].

Because we focus on the polymerization process of biopolymers, we assume the existence of sufficient amounts of (activated) monomers and do not discuss the synthesis of monomers.

Synthesis of longer biopolymers like peptides or nucleotides is another problem. For example, the Gibbs energy for peptide bond formation between glycine molecules is $\sim 6k_BT$ (at 25 °C and neutral pH) [24, 25]; therefore, the yield of the dimer is $\sim 0.2\%$. Similarly, for the phosphodiester bond, this energy is $\sim 9k_BT$ [26], and thus the yield of the dimer is $\sim 0.01\%$. (On the other hand, if the monomer of nucleotides is activated, formation of the bond is energetically favorable.) There are various experiments involving nonenzymatic polymerization of peptides and nucleotides on the prebiotic Earth. For example, in the experiment that uses mineral clay as a catalyst, RNA like polymers up to 14-mer are synthesized from activated monomers [27, 28]. Among different kinds of biopolymers, their length in vitro experiments on polymerization obeys the Flory–Schultz distribution [29] $f(l) = a^2 l(1-a)^{l-1}$, where l is the length of polymers, and a is a constant. This distribution is derived by the process in which a polymer elongates at one of the termini at rate $1-a$ and stops its elongation at rate a. This exponential decay with certain length is natural because longer polymers have a higher chemical potential.

1.1.2 Experimental construction of a self-replicating system of polymers

Using the Qβ replicase ribozyme, Spiegelman has attempted to evolve RNA sequences experimentally [30, 31, 32] (This is a experimental implementation of the quasispecies model discussed in section 1.2.1). As a consequence of this evolution, a "little monster" of smaller length appears that can replicate faster than the initial longer RNA polymer. The method has been extended to directed evolution of enzymes and is used for enzymes of interest in bioengineering [33].

Such evolution experiments require other enzymes or ribozymes; therefore, the system itself is not autocatalytic. An autonomously replicating system consisting only of RNA is implied in the hypothesis about the origins of life such as the RNA

world. Nevertheless, in the present natural world, such examples of life forms are not known. Accordingly, researchers have attempted to construct such a system in a laboratory. For example, ribozymes that catalyze the polymerization of two fragments of itself have been synthesized by Joyce's group [34]. In this system, a ribozyme can replicate exponentially. Thus, Darwinian evolution (selection of the fittest) is observed among different kinds of fragment resources. Alternatively, Lehman's group [35] has constructed a network of mutually catalyzing ribozymes. Each ribozyme can be formed by binding of several parts and catalyze the formation of covalent bonds of others.

On the other hand, the synthesis of the RNA polymerase ribozyme (RPR) has also been attempted by Joyce and Holliger [36, 37, 38, 39]. RPR is a ribozyme that catalyzes the reaction of RNA replication. As the highest record to date, they succeeded in replicating RNA as long as 200 bases. This result does not reach the length of RPR . Holliger and Joyce created RPR with a catalytic ability by complex formation of noncatalytically synthesizable fragments as big as 30 bases.

1.2 Theoretical studies on the origins of life

Self-replication is one of the most fundamental properties of life. Here, we first review Eigen's theory of such self-replication [40, 41]. Next, we will examine the studies on the emergence of such an autocatalytic system. Finally, we will review the chemical evolution toward a more complex protocell.

1.2.1 Eigen's error catastrophe and the hypercycle (selection and evolution law for self-replicating templates)

Phenomenological theory of growth and selection

We start from a simple chemical reaction in a container of constant volume:

$$(S) + X \xrightarrow{r} 2X.$$

In this reaction, a molecule (replicator), X, produces a copy of itself at rate r while consuming substrate (nutrient) S, which is assumed to be present in a constant amount. Here, we postulate that external nutrients are supplied rapidly. Hence, the concentration of S is assumed to be constant.

The corresponding rate equation is of course

$$\dot{x} = rsx,$$

where s and x are concentrations of S and X, respectively. (We will set s to 1 for simplicity.) Solution by elementary calculation leads to $x(t) = x_0 \exp(rt)$.

Instead of the growth in proportion of the concentration itself, another form of growth is sometimes discussed:

$$\dot{x} = rx^{\alpha}.$$

(Examples of this type will be provided later.) Depending on the exponent of growth rate α, the growth of the replicator is different qualitatively. When $\alpha < 1$, the replicator grows subexponentially (or parabolically), whereas at $\alpha > 1$, the growth is superexponential (or hyperbolic). The differential equation can be solved as follows: $x(t) = \left[x_0^{1-\alpha} + (1-\alpha)rt\right]^{\frac{1}{1-\alpha}}$ ($\alpha \neq 1$). Especially, for $\alpha > 1$, the concentration of the replicator goes to infinity at $t = x_0^{1-\alpha}/(\alpha - 1)r$.

Next, we consider multiple replicators $X_1, X_2, \ldots, X_N$ and competition among them. We introduce their dilution so that the total sum of their concentrations is constant (i.e., $\sum_i x_i = x_{total} = const$) [2] In other words, the replicators share the same "niche" (competition for the same resources).

$$\dot{x}_i = r_i x_i^\alpha - \phi x_i,$$

where $\phi = \frac{1}{x_{total}} \sum_j r_j x_j^\alpha$. (We set $x_{total} = 1$ later.)

The stationary states (fixed points) of the above rate equation qualitatively change depending on exponent α. When $\alpha = 1$ (the case of exponential growth),

[2]There are several ways to implement the stationary growth of replicators. (without diverging the concentrations to infinity), for example, the chemostat condition (proposed by Leo Szilard [42]) or a simultaneous increase in container volume. In the chemostat condition, a constant supply of resource S (at rate f) and dilution of all the compounds (at rate d) are considered.

$$\dot{s} = f - s\sum_i r_i x_i - ds$$
$$\dot{x}_i = s r_i x_i - d x_i$$

In this equation, total concentration $s + \sum_i x_i = f/d$ is conserved. s can be eliminated as follows: $s = \frac{f}{d + \sum_i r_i x_i}$ ($\dot{s} = 0$) or equivalently $s = f/d - \sum_i x_i$.

Another possibility is reaction systems in a protocell coupled with vesicle growth. If the number of replicators X_i, n_i, obeys the rate equation

$$\dot{n}_i = s_0 r_i n_i^\alpha / V^{\alpha - 1},$$

then the rate equation of concentration $x_i = \frac{n_i}{V}$ is

$$\frac{d}{dt}\left(\frac{n_i}{V}\right) = \frac{\dot{n}_i}{V} - \frac{\dot{V} n_i}{V^2}.$$

If we let the volume V grow proportionally to the total number of replicators $\sum_i n_i$, then $\dot{V}/V$ corresponds to dilution rate ϕ in the text.

Although the two equations have different dynamics (for instance, in the chemostat condition, there is only a fixed amount of monomers), the stationary state is equivalent when we set d to $\phi^*, f = d(s_0 + x_{total})$, where ϕ^* is ϕ in the stationary state.

Note that rescaling of concentration ($x_i/x_{total} \to x_i$ and $x_{total}^\alpha \to t$) is incorrect when there is a different order of terms in the rate equation. For instance, let us consider rescaling of the rate equation,

$$\frac{d}{d(x_{total} t)} = r/x_{total}(x/x_{total}) + r'(x/x_{total})^2.$$

If x_{total} is infinite, the first term goes toward zero. Therefore, in the size limit of a large system, the only effective term in the rate equation is the one with the highest order. For this reason, the hypercycle can be dominant in the system.

the only (nontrivial) stable fixed point is the one corresponding to the survival of a replicator with the highest replication rate ($x_i = 1(i = a), 0(i \neq a), a = \arg\max_j r_j$). That is, if the replication rates are given as fitness, this means "survival of the fittest." (If those replicators have exactly the same maximal replication rate ($r_i = r_j$), then their concentrations are indeterminate.) For $\alpha < 1$ (subexponential growth), at a stable fixed point, all the concentrations of a replicator have a finite value: i.e., all the replicators coexist. The solution is $x_i = \frac{[(1-\alpha)r_i]^{\frac{1}{1-\alpha}}}{\sum_j [(1-\alpha)r_j]^{\frac{1}{1-\alpha}}}$. At $\alpha > 1$ (superexponential growth), only one type of replicator survives, and this type is dependent on the initial state of replicators' concentrations. ($x_i = 1$ and $x_j = 0(i \neq j)$.) The one that is present in a larger amount can grow faster and tends to be selected (frequency-dependent selection).

The problem of error catastrophe in the quasispecies model

We consider bit sequences of length L, $\{0,1\}^L$ as templates. During the replication process, an error occurs at each bit with constant probability μ. (See the section on kinetic proofreading regarding the physics behind the determination of μ.)

The rate equation is

$$\dot{x}_i = \sum_j w_{ij} x_j^\alpha - \phi x_i,$$

where $\phi = \sum_{i,j} w_{ij} x_j^\alpha$. And

$$w_{ij} = r_j \binom{L}{h_{ij}} \mu^{h_{ij}} (1-\mu)^{L-h_{ij}},$$

where h_{ij} is the Hamming distance between bit sequences i and j.

We designate the 0th sequence (the sequence with all zeros) as a master sequence. We suppose that $\alpha = 1$ and the replication rate of the master template is faster than that of others. Therefore, only the master sequence survives as discussed in the previous subsection. The question is whether the information of the master sequence persists under the conditions of a nonzero mutation rate.

For simplicity, we assume that the replication rate of the master sequence is r_0, and these rates of the others are r_1. We define order parameter $\Phi = \sum_i (1 - 2h_{0i}/L)x_i$, where h_{0i} is the Hamming distance between the master sequence and the ith sequence. This parameter means the accuracy of the information of the master sequence or average similarity of the templates to the master sequence. If only the master sequence exists, then $\Phi = 1$. If the distribution of template concentrations is uniform, then $\Phi = 0$. As presented in Fig. 1.1, the fraction of the master sequence and the order parameter decrease with an increase in mutation rate μ. Above a critical value, the master sequence cannot persist because the number of incorrect sequences exponentially increases with L.

Judging by the condition for the growth of the master sequence, the latter is defeated by the growth of other sequences:

$$r_0(1-\mu)^L \sim r_1,$$

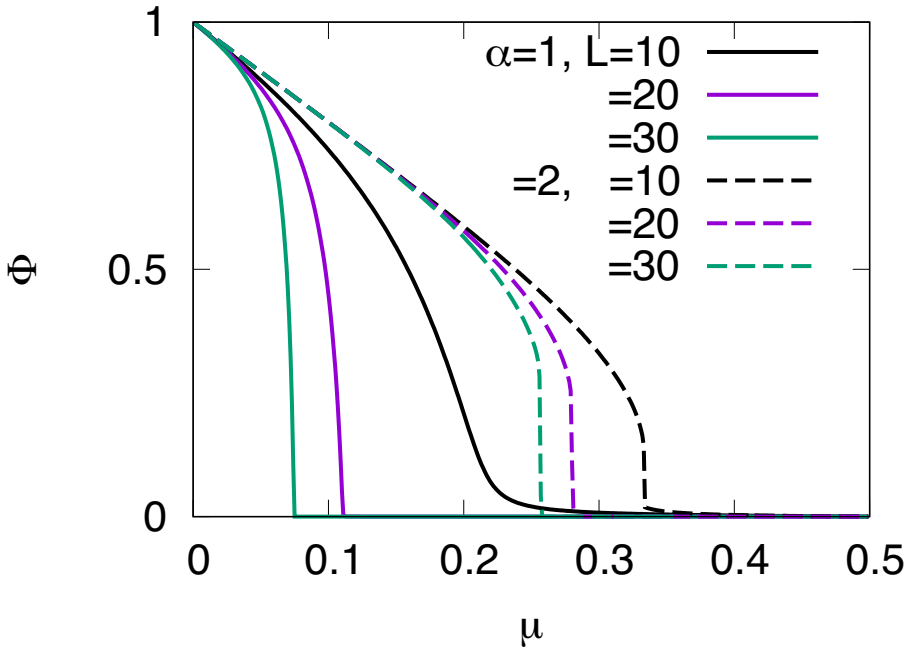

Figure 1.1. Dependence of the order parameter Φ on the error rate μ for different α and L. Φ is defined in the main text.

and critical mutation rate μ^* is

$$\mu^* \sim \frac{1}{L} \ln(r_0/r_1).$$

From the rough estimation that $\mu \sim 10^{-2}$, Eigen concluded that in this case of exponential growth, replication of information of sufficient length ($L \gtrsim 100$) is not plausible. On the other hand, if α is over 1 (if the growth of templates is superexponential), then the critical error rate drastically increases, although the critical rate still decreases with the increase of template length. This is because the master sequence, which is in the majority under the initial conditions, has an advantage during frequency-dependent selection. Therefore, in the prebiotic world, to replicate information of sufficient length, a mechanism is needed that makes α greater than 1.

On the other hand, at $\alpha < 1$, the error threshold is even worse than in the exponential-growth case. This is because even at $\mu = 0$, wrong sequences coexist that have lower fitness than the master sequence does.

The competition between the difference in fitness and the number of sequences is quite similar to the relation between energy and entropy in thermodynamics and statistical physics. Error catastrophe is described as the first-order phase transition in a statistical physics model of replicator equation [43, 44, 45]. (Actually, μ increases with the rising temperature.) Besides, the general case of the fitness landscape has been studied.

Implementation of superexponential growth: the hypercycle ($\alpha > 1$)

The growth rate exponent of more than 1 is interpreted as cooperation among replicators:

$$(S) + X + X \to 2X + X$$

In this replication reaction, more than one replicator is involved.

Eigen suggested a cooperative network of replicators as the mechanism underlying their superexponential growth,

$$X_i + X_j \stackrel{c_{ij}}{\to} 2X_i + X_j.$$

In the network, a replicator catalyzes a replication reaction of another replicator. The corresponding rate equation is

$$\dot{x}_i = \sum_j c_{ij} x_i x_j - \phi x_i,$$

where $\phi = \sum_{i,j} c_{ij} x_i x_j$.

For instance, the cycle of n components $X_i + X_{i+1} \to 2X_i + X_{i+1}$ ($X_{n+1} = X_1$), their total concentration $x_{cycle} = \sum_i^n x_i$ grows as follows: $\dot{x}_{cycle} = x_{cycle}^2/n$. Thus exponent α is 2. [3]

Nonetheless, in the hypercycle, the problem of parasites could emerge. Suppose a template catalyzes replication of itself and of another, but the latter molecule does not catalyze any reaction:

$$\dot{x}_1 = c_{12} x_1 x_2 - \phi x_1,$$
$$\dot{x}_2 = c_{22} x_2^2 - \phi x_2.$$

(If time t is rescaled as tx_2, then the replicator's growth is exponential.) In this case, the selection is the same as in the case $\alpha = 1$; thus, if $c_{12} > c_{22}$, eventually the parasite will dominate the whole population, and replicators will not be able to continue to replicate. (In this case, X_2 is considered a polymerase ribozyme, and X_1 is a shorter one that lost catalytic activity. It is only logical that the shorter template can replicate faster just as Spiegelman's "little monster" can). The solution to this problem is discussed in detail later in the section about multilevel selection.

Note that the hypercycle is different from cross-catalysis in which the growth rate exponent equals unity [47]. ($X_1 \to X_2 + X_1, \quad X_2 \to X_1 + X_2$.) Although experimental implementations of cross-catalysis were reported recently [35], this approach has not succeeded yet in the construction of the hypercycle.

---

[3] More generally,

$$\dot{x}_i = f_i(\{x_j\}) x_i - \phi x_i,$$

where $\phi = \sum_i f_i(\{x_j\}) x_i$.

The fitness (replication rate) function is determined by concentrations of all the replicators. This type of equation is studied mostly in evolutionary game theory [46]. The hypercycle with a parasite corresponds to the prisoners' dilemma game.

Another possible way to implement the superexponential growth of a replicator is multistep autocatalysis. For example,

$$(S) + X \to X' + X \to X + X, \quad X' \xrightarrow{d} (S).$$

In this replication reaction, X is produced by passing through intermediate state X'. In addition, there is a backward reaction where X' goes back to S or is diluted from the system. This type of chain reaction is realized as stepwise ligation reactions (polymerization) of polymers [48] or a mutually catalytic metabolic network [49]. The rate equation is

$$\dot{x}' = x - x'x - dx', \quad \dot{x} = x'x.$$

When $d \gg 1$, this equation can be reduced to

$$\dot{x} = \frac{1}{d}x^2,$$

thus corresponding to the case of $\alpha = 2$ in Eigen's model of growth.

Implementation of subexponential growth ($\alpha < 1$)

The replication of RNA templates in the experiment by Kiedrowski [50] shows subexponential growth. To be precise, in the replication of a nucleotide sequence, a base pair is involved [51]. Therefore, the double helix of templates must be dissociated to start another replication process.

$$(S) + X \xrightarrow{r} C, \quad C \underset{k_f}{\overset{k_b}{\rightleftarrows}} 2X,$$

where X is a free template, and C is a double helix (complex) of a pair of templates. The corresponding rate equation is

$$\dot{x} = 2k_b c - 2k_f x^2 - rx,$$
$$\dot{c} = -k_b c + k_f x^2 + rx.$$

In this case, the growth of templates becomes asymptotically subexponential. We assume that $r \ll k_b \ll k_f$(dissociation of templates is energetically difficult). Therefore, from conservation of the total concentration of a template, it follows that $x_t = 2c + x \sim 2c$ (because $k_b \ll k_f$). From $k_b c = k_f x^2$, it can be concluded that

$$\dot{x}_t = r\sqrt{\frac{k_f}{2k_b}x_t}.$$

This is corresponding to the case of $\alpha = 1/2$ in Eigen's model of growth.

The finite-population case

So far, we have considered only the model with a deterministic and continuous rate equation. By contrast, in the evolution and selection of a replicator, stochasticity could be an important factor. Moreover, discreteness of the number of molecules is essential for describing the phenomena of biopolymers because the combination of sequences is far larger than the number of template polymers.

In population genetics, the stochastic property of evolution of a replicator is studied. The simplest model of replicators with a finite population is the Moran process [52, 53]. Let there be two types of replicators X and Y, and their number is always fixed at N. When a replicator makes its copy, one replicator is randomly chosen and removed from the system to fix N. We set the number of X species to n. Then,

$$\tau_+(n) = r_X(1-\mu)(1-n/N)n + r_Y\mu(1-n/N)(N-n),$$
$$\tau_-(n) = r_Y(1-\mu)(n/N)(N-n) + r_X\mu(n/N)n.$$

are transition probabilities from state n to $n+1$ and from n to $n-1$, respectively. r_X and r_Y are replication rates of X and Y, and μ is a small rate of mutation at each replication. Then the stochastic dynamics of this chemical reaction system follow the master equation,

$$\dot{P}(n,t) = \tau_+(n-1)P(n-1,t) + \tau_-(n+1)P(n+1,t) - (\tau_+(n)+\tau_-(n))P(n,t).$$

In the large N limit, if we set x to n/N, the master equation becomes a simple rate equation $\dot{x} = r_X(1-\mu)x + r_Y\mu y - \phi x$, $\dot{y} = r_Y(1-\mu)y + r_X\mu x - \phi y$, where $\phi = r_X x + r_Y y$ (corresponding to a replicator equation with mutation $L = 1$).

Let us consider the case where one X molecule appears in a Y-dominated population. We can calculate the probability of fixation of the X molecule (the probability that X overtakes Y),

$$\rho = \frac{1-\frac{r_Y}{r_X}}{1-(\frac{r_Y}{r_X})^N}.$$

Then, the speed of evolution v is given by

$$v = N\rho\mu$$

If the replication rates of two replicators are identical ($r_X = r_Y$), $\rho = 1/N$; thus, v does not depend on system size N. For this reason, the number of mutations in genetic information can serve as a "molecular clock," i.e., a method to measure time because of differentiation of species.

Via Kramers–Moyal expansion [54, 55], the master equation is approximated by the Langevin equation with continuous variable $x = n/N$:

$$\dot{x} = r_X\mu(1-2x) + \sqrt{\frac{r_X}{N}\left[\mu + 2(1-2\mu)x - 2(1-2\mu)x^2\right]}\eta(t)$$

where $\eta(t)$ is a random variable with $\langle \eta(t)\eta(t') \rangle = \delta(t-t')$ Then, in case of $\mu > 0$, the stationary-state distribution of x can be calculated as

$$P^{st}(x) \propto (\mu + 2(1-2\mu)x - 2(1-2\mu)x^2)^{\gamma}$$

where $\gamma = \frac{N\mu}{1-2\mu} - 1$[4].

In the limit of infinite system size, the stable fixed point is $x = 1/2$, i.e., replicators coexist. When N is large, in the probability distribution, there is a single peak near the fixed point. On the other hand, when N is small, there are two peaks near $x = 0, 1$. This is because when the number of replicators is small, the time to extinction (or fixation) is short. (Regarding the chemical Langevin equation, multiplicative noise is significant at $x = 1/2$ and small at $x = 0$ or 1. Therefore, the system stays for a long time at the edge of x.)

This phenomenon is similar to discreteness-induced transition in the Togashi–Kaneko model [56]. The effect of smallness of noise in the chemical reaction system has extensively been studied [57, 58, 59, 60, 61]. Especially, homochirality of biopolymers is explained by a similar model [62].

Next, we consider the effect of noise to the problem of error catastrophe. When there are not only mutational errors but also stochastic noise due to the finiteness of the population, error tolerance of the master sequence is even worse [63]. This effect is related to Muller's ratchet. When the population of the replicator is finite, as in the above discussion, even in the case of the dominance of the master sequence, there is a chance of fixation of a replicator with lower fitness.

In case that replicators grow with the exponent $\alpha > 1$, there are multiple fixed points in which each replicator dominates as discussed previously. In contrast to infinite system size and deterministic case, the transition between attractors is mediated by noise due to the finiteness of system size. If the dynamics are rephrased by the flow in the potential, an attractor representing a faster replication rate has a deeper well in the potential landscape. Therefore, in a stationary distribution, the fittest is selected even if the growth of replicators is superexponential.

More generally, when the growth rate of X_i is a function of Xs, i.e.,

$$\dot{x}_i = f_i(\{x_j\}_j) - \phi x_i.$$

and there are multiple attractors, then the attractor with the fastest growth ϕ^* is also selected by stochastic noise. Because in principle, strength of attraction to a stable fixed point (maximum eigenvalue of a Jacobian) is $-\phi^*$ [5], and the depth of the potential valley for a given attractor positively correlates with the attraction. This attractor selection mechanism is employed by living systems such as bacteria for adaptation to different environments [64].

[4]The solution for the Fokker–Plank equation $\partial_t P = \partial_x(A(x)P) + \frac{1}{2}\partial_x^2(B(x)P)$ is $P = \frac{1}{B(x)} \exp(\int^x \frac{A(x')}{B(x')}dx')$.

[5]If $\partial_j f_i(x^*) \sim 0$ is assumed to be near fixed point x^*, Jacobian $J = \{\partial_j f_i(x^*) - (\sum_k \partial_j f_k(x^*)x_i^*) - \phi^*\delta_{ij}\}_{ij} \sim \{-\phi^*\delta_{ij}\}_{ij}$, thus $|J| = -\phi^*$

The thermodynamics of replicators

Normally, the reaction that converts substrate S into replicator X never occurs (or is very rare) because X is thermodynamically less stable than a substrate.

In a replication reaction, X itself determines the rate of the reaction (that it catalyzes) drastically. On the other hand, the existence of a catalyst does not change the thermal equilibrium state. Therefore, the supply of energy, i.e., a nonequilibrium condition is required for the replication reaction to proceed.

Consequently, in an autocatalytic system of a nonequilibrium steady state, usually, energy is provided via a supply of nutrients S and is dissipated by removal of replicators X.

For example, if the replication of replicator X is a reversible reaction,

$$S + X \underset{k^-}{\overset{k^+}{\rightleftarrows}} 2X,$$

then entropy production is expressed as

$$\sigma = x(k^+ s - k^- x)\mathcal{A}/T,$$

where $\mathcal{A}$ is the chemical potential difference between S and X $\mathcal{A} = ln|k^+ s| - ln|k^- x|$, and T is the temperature of the system. To maintain this autocatalytic system, the same amount of energy supply is required [65].

1.2.2 Emergence of autocatalytic sets

In the previous section, we considered a phenomenological theory for replicating polymers (an autocatalytic set). Nevertheless, how such replicators or their system have emerged in the first place? This question is what we want to discuss in this section.

Generally, such replicating polymers do not emerge spontaneously because such long and complex polymers are rather energetically unstable. Because reactions of their synthesis take place only rarely, enzymes or ribozymes are necessary to speed up the synthesis reactions.

Pioneering work for the study of the emergence of autocatalytic set has been done by Freeman Dyson [66]. Suppose there are polymers of length L that contains $m+1$ monomer species: a correct and energetically active one and m incorrect and inactive ones. Monomers have a chance to change their type with a mutation. A polymer composed of active monomers has a catalytic ability that can help inactive monomers to gain activity. He assumed that the activity of the polymer is $\psi(\xi)$, where $\xi(= n/L)$ is the proportion of active monomers in it. From rough thermodynamic argument, he decided that $\psi(\xi) = \exp(U\xi/k_B T)$, where U is the difference of the activation energy at the presence of a perfect catalyst. Then, transition probabilities from state with n to $n+1$ active monomers and from n to $n-1$ are

$$\tau_+(n) = \frac{\psi(\xi)}{m+1}(1 - n/L), \qquad \tau_-(n) = \frac{m}{m+1} n/L.$$

Then, the rate equation corresponding to this stochastic process,

$$\dot{\xi} = \psi(\xi)(1-\xi) - m\xi,$$

has two stable fixed points: a high activity state and low activity state.

The model is mathematically the same as the Ising model in a fully connected graph. The transition between inactive and active states is regarded as a first-order phase transition in statistical physics. If the system size is finite, the transition between states occurs due to stochastic noise like the transition between ordered states in the Ising model. Just as in the Ising model, the transition probability is calculated approximately in the Arrhenius law–like form (or Kramers rate), i.e., an exponential of energy barrier $\propto \exp(L\Delta E)$, where $1/L$ is corresponding to the strength of noise.

Within this framework, Dyson tested whether the RNA world or the protein world is more plausible as the first autocatalytic set. Using the phenomenological parameters m, L and U for proteins and nucleotides, he calculated the transition probability from the inactive to active states. If this probability is greater, one can conclude that the emergence of the autocatalytic set is more plausible. With his conclusion, the protein world (autocatalytic set with protein) is more plausible as prelife.

In Dyson's phenomenological model, the concept of polymerization is lacking. Not only thermodynamics but also kinetics are vital for estimating the transition probability because it is a dynamical property.

$$\dot{x} = (1-x)(\epsilon + x^{\alpha}) - dx$$

x is concentration of a "catalyst," and $(1-x)$ is concentration of "monomers." Polymers are dissipated from the system or decay back into monomers at rate d. A catalyst synthesizes itself from a monomer at a unit rate but is synthesized spontaneously at (small) rate ϵ. For exponential growth rate $\alpha > 1$, a cooperative structure of the catalyst is required (see the section about the hypercycle).

More specific polymer models are also studied. In most of such studies, concrete polymers with certain length l of multiple types χ of monomer sequence $\{\chi\}^l$ are considered a ligation reaction between polymers and a cleavage reaction of a polymer:

$$A + B \to AB, \qquad AB \to A + B$$

where A and B are polymers with arbitrary length and sequence and these reactions are catalyzed by another polymer C with catalytic efficiency k. Usually, only monomers (or short polymers) are supplied, and for the synthesis of a specific polymer, several steps (reactions) are required. For example, to synthesize a 2^n-mer from monomers requires at least n steps of ligation reactions. Thus, in contrast to Dyson's model, the synthesis of a catalyst is constrained not only energetically but also kinetically.

For instance, in the model by Kauffman et al. [67, 68], it is assumed that randomly chosen polymer species are the catalyst. That is, a reaction among substrate

A and B and catalyst C exists at probability ρ. Thus, polymers constitute a network of mutual catalysis at connection rate ρ. As a conclusion of the model, at critical ρ^*, the network is percolated, and new longer polymers are synthesized infinitely (if resources are unlimited).

Giri et al. [69] have stated how strong a catalyst needs to be for an autocatalytic set to emerge if we assume dissipation of polymers in the system. They not only have to be connected to each other but also the catalyst should be sufficiently catalytically active. There is a threshold for the catalytic efficiency of a catalyst for an autocatalytic set with longer polymers. Nonetheless, the threshold can be reduced if there is an autocatalytic system of shorter polymers with smaller catalytic activity.

Guseva et al. [70] analyzed the mechanism of action of the catalyst concretely. As monomers, hydrophobic H and polar P are assumed, and the folded structure is calculated via the scheme of the HP model [71]. By rough thermodynamical estimation, the rate of a ligation reaction between polymers via condensation at a hydrophobic surface of the folded polymer is derived. They concluded that although this catalytic efficiency is small, it is enough for the appearance of autocatalytic set with long polymers.

On the other hand, there is a series of studies in which template-directed ligation is specifically considered as a catalyzed reaction. That is, a ligation between A and B polymers occurs via forming base pairs with a sequence of another polymer C.

For example, the symmetry breaking in the polymer distribution and the emergence of information is discussed in studies of such a model. In Anderson's model [72, 73], polymers are ligated in a template-directed way. Besides, the death rate of polymers depends on their sequences and is determined as quenched random values as in a spin glass model. As a result of the model, the system manifests multiple stable fixed points with different polymer distributions.

In a realistic model that details the dynamics of an in vitro experiment on template polymerization by Kiedrowski [50], selection of a specific motif of the sequence is observed [74]. Similarly, Tanaka et al. have reported that in a simplified model of a template-directed ligation system, only periodic sequence is observed [75].

Obermayer et al. [76] considered the stability (free energy) of a folded structure of a duplex of two nucleotides, and the cleavage rate depends on it (i.e., a reaction $AB \to A+B$ is inhibited by the presence of the C polymer which forms a duplex with the AB polymer). This approach plays a similar role in template-directed ligation during the production of sequences, and information inheritability is observed in the model.

In the model of Ohtsuki et al. [77, 78], there are not only ligation reactions between polymers ($A+B \to AB$) but also phenomenological replication of a sequence ($A \to 2A$). As in a quasispecies model, if the replication rate is fast enough and the mutation rate is low enough, information inheritance is possible.

Besides, in the models by Tkachenko et al. [79] and by Wu et al. [80], they discussed the emergence of long polymers via template directed ligation. They considered template-directed ligation in a mean field likely for base pairs between

sequences. Then, they showed that the process could produce an autocatalytic set of longer sequences.

Multiple types and a limitation of resources (niche) and complexity

So far, we have considered only the case of only one type of resource (or niche). In such a case, (except for the case with the subexponential growth of replicators), only one replicator can survive in a stationary state. On the other hand, in a situation involving several types of resources, several species of replicators can survive. This is because if different species consume different resources, then there is no competition (interaction) between them. Generally, the number of survivable species of replicators does not exceed the number of different resources or niches; this is known as the Gause's principle [81, 82].

In this regard, let us examine the case where each replicator X_i's growth depends on the logistic equation in the same niche (same limited space or requirement for growth of a shared resource),

$$\dot{x}_i = r_i(1 - x_i/K_i)x_i - \phi x_i,$$

where $\phi = \frac{1}{x_{total}} \sum_i r_i(1 - x_i/K_i)x_i$. K_i is capacity of X_i, which is a stationary concentration without competition with other replicators.

The consequence of selection depends on x_{total}, which can be seen as "strength of selection pressure." In the case of small selection pressure, all the species coexist, and in the opposite case, the species with the highest fitness survive (survival of the fittest). (For example, in a two-species case and if we assume that $r_1 > r_2$, the threshold is $x_{total} = \frac{K_1}{r_1}(r_1 - r_2)$.) This difference in selection is known as r/K selection theory in ecology [83].

For a more complicated autocatalytic set, maintaining the diversity of components in an autocatalytic system is essential. For example, diversification of components is discussed in the study on the hypercycle [84]. Here, we consider the hypercycle with N components, where each component X_i consumes a different substrate, S_i, during its replication catalyzed by X_{i+1};

$$S_i + X_i + X_{i+1} \xrightarrow{1} 2X_i + X_{i+1}, \quad \emptyset \underset{D}{\overset{D}{\rightleftharpoons}} S_i,$$

where $X_{N+1} = X_1$. Here, substrate S_i diffuses in and out at the rate D.

The corresponding rate equation for the reaction system is

$$\begin{aligned} \dot{s}_i &= D - s_i x_i x_{i+1} - D s_i, \\ \dot{x}_i &= s_i x_i x_{i+1} - \phi x_i, \end{aligned}$$

where x_i and s_i are concentrations of X_i and S_i, respectively. Here, we suppose that total concentration of Xs is 1 ($\sum_i x_i = 1$). Furthermore, then if we assume that all the Xs are homogeneous, then the concentration of one of them is $x_i = 1/N$.

Thus, in the stationary state, growth rate ϕ^* is derived as

$$\phi^* = \sum_i s_i x_i x_{i+1} = \frac{DN}{1 + DN^2}.$$

Then the rate is maximal at $N = 1/\sqrt{D}$. Therefore, for smaller diffusion rate D (i.e., when substrates are depleted), a hypercycle with more components is advantageous with respect to the growth rate.

1.2.3 A compartment and multilevel selection

Error correction by compartmentalization

Until now, we have analyzed the chemical reaction system of replicators under conditions of good mixing, and there is no spatial structure. One of the properties that life must have is compartmentalization. Modern cells have a double lipid layer, which isolates the internal chemical reactions from those outside. Similarly, in the origins of life, at a certain point of chemical evolution, a prelife autocatalytic system has to become a compartment system.

Compartmentalization is one of the ways for the exclusion of parasites or for error correction in components of an autocatalytic set [85, 86]. When there is a small population in one compartment, there is some probability that a compartment where there are few or no parasites could appear, and it can prosper because it has more of the catalyst and grows faster. A similar effect is also proposed in ecology or sociology as the exclusion of a selfish individual or a free-rider (group selection).

To discuss the stochastic corrector mechanism, here, we review the parasite problem in the hypercycle with a finite population and a multilevel Moran process. We suppose that there are replicators of N fixed population competing for the same niche in a compartment. In addition, there are M such compartments, and their interior is separated spatially from that of other compartments.

The master equation of such a system is

$$\begin{aligned}\dot{P}_m(n) = \tau^+(n-1)P_m(n-1) + \tau_m^-(n+1)P_m(n+1) - P_m(n)(\tau_m^+(n) + \tau_m^-(n)) \\ + \frac{1}{M-1}\sum_{m' \neq m}\left[P_{m'}(n)\tau_{m'}(n) - \sum_{n' \neq n} P_{m'}(n')\tau_{m'}(n')\right],\end{aligned}$$

where the equation in the first line represents a Moran process at the inner compartment level, and the second line represents the process at the between-compartment level. $P_m(n)$ is the probability that the mth compartment contains n catalytic replicators from the total of N replicators. $\tau_m^+(n)$ and $\tau_m^+(n)$ are transition probabilities from n to $n+1$ and n to $n-1$ in mth compartment, respectively. $\tau_m(n)$ is the probability of the mth compartment with n catalytic replicators to divide into two compartments: $\tau_m(n) = n$.

From the stochastic simulation of the master equation, for small N and large M, catalytic replicators prosper relative to the whole population in the stationary state. In the opposite case, parasitic replicators prosper.

In the same way, this error correction mechanism works for the cases with the exponential growth of replicators, and for one of the solutions to prevent Muller's ratchet, which is discussed above.

Origin of genetic system

Other than compartmentalization, the prelife system has to acquire a complex biological mechanism. One of such "major transitions" [87] toward LUCA is acquisition of the genetic mechanism. In modern cell systems, one or a few DNAs are transcribed into mRNAs, and mRNA is translated into enzymes. Why did life invent such a mechanism or what is the benefit for a living system to have it? One of the possible answers is that for a whole cell to have an "evolvability" property, it is necessary that one of the replicator molecules be in the minority and undergo strong selection [88].

Consider the system in which two types of replicators X and Y catalyze each others' replication.

$$X_i + Y_0 \to X_i + X_j + Y_0, \quad Y_i + X_0 \to Y_i + Y_j + X_0$$

The replication can be catalyzed only by catalytically active master sequence X_0, Y_0, and errors occur during the replication at constant rate μ.

The rate equation of this reaction's dynamics is

$$\begin{aligned} \dot{x}_i &= r_X \sum_j w_{ij} x_j y_0 - \phi x_i \\ \dot{y}_i &= r_Y \sum_j w_{ij} y_j x_0 - \phi y_i \end{aligned} \tag{1.1}$$

where $\phi = (r_X \alpha_x + r_Y \alpha_y) x_{tot} y_{tot}$. ($x_{tot} = \sum_i x_i$, $y_{tot} = \sum_i y_i$, $\alpha_x = x_0/x_{tot}$, $\alpha_y = y_0/y_{tot}$.)

Same as in the quasispecies model, sequence information of X and Y suffers error catastrophe because there is no fitness difference between an active and inactive one if we have a large population and a deterministic case. Here, we presume a finite population of Xs and Ys in a compartment, and competition between them as before. We also assume that $r_X \gg r_Y$, i.e., production of Y is much slower than that of X, and thus Y is in the minority in a compartment. In this case, via a stochastic corrector mechanism (as discussed above), strong selection pressure leads to a greater proportion of active Y molecules Y_0 in a protocell.

Minority molecule Y (in the amount of Y_0) determines the growth speed of the compartment. Therefore, it has the same role as a genetic polymer, just as DNA has for modern cells.

1.3 Kinetics and thermodynamics of polymerization

The polymerization reactions of biopolymers such as RNA, DNA, or proteins are the fundamental processes in living systems. The kinetic and thermodynamic properties of the polymerization are studied separately from the context of the origins of life.

In a copolymerization process of DNA or RNA in vivo, various enzymes including polymerases are involved. A simplistic picture of such a process is energetically driven one-dimensional random walk of a polymerase on a template polymer. Under this assumption, kinetics of the polymerization have been studied. The dynamics of the probability that the polymer is of the length l and the monomer conformation w, $P(l, w)$ is described as

$$\dot{P}(l, w) = \tau_- P(l + 1, w'') + \tau_+ P(l - 1, w') - (\tau_+ + \tau_-)P(l, w).$$

For example, if we hypothesize copolymerization of a random nucleotide sequence and that the velocity of replicases depends on the types of bases at the site, the speed of polymer elongation is sublinear (i.e., $\langle l \rangle = t^\alpha$, $0 \leq \alpha \leq 1$) [89]

Furthermore, the thermodynamic property of polymerization is analyzed [90, 91]. As a consequence of the second law of thermodynamics ($\dot{S_{tot}} \geq 0$),

$$\frac{dS_{tot}}{dt} = vA \geq 0$$

is derived[6], where $v = \langle \dot{l} \rangle$, $A = \epsilon + D(w) + I(w, \alpha)$. ϵ is driving energy, $D(w)$ is Shannon entropy of the polymer sequence, and $I(w, \alpha)$ is mutual information between the polymer sequence and template sequence. This condition restricts the minimum energy supplied to the proceeding polymerization reactions.

Each monomer addition during the polymerization can be regarded as a Michaelis–Menten type of reaction:

$$R/W + E \underset{k^-_{r/w}}{\overset{k^+}{\rightleftarrows}} R/W \cdot E \xrightarrow{k} P_{r/w}.$$

Suppose the energy of complex formation between the right or wrong substrate R/W (monomer) and catalyst E (template) is e_r and e_w, respectively. In addition, we set the association rate of both substrates to 1 for simplicity. Accordingly, the dissociation rates of both are $k^-_r = e^{\beta e_r}$ and $k^-_w = e^{\beta e_w}$.

The production rates of correct and wrong products are (assuming that the amount of the catalyst is limited),

$$\dot{p}_{r/w} = \frac{k x_{r/w} e}{1/k^-_{r/w} + x_{r/w}}.$$

[6]This equation is obtained via a time derivative of $S_{tot} = -\sum_{l, w \in l} P(l, w) \log(P(l, w))$ using the ansatz $P(l, w) = P(l)P(w)$ [90].

Therefore, when concentrations of substrates x_r and x_r are identical and small, the proportion of the production rate of the wrong one is

$$\mu = \frac{\dot{p}_w}{\dot{p}_r + \dot{p}_w} \sim \frac{e^{-\beta e_w}}{e^{-\beta e_r} + e^{-\beta e_w}} \sim e^{\beta(e_r - e_w)}.$$

For instance, these reactions are regarded as the addition of one monomer during the polymerization process. E is the template, and R or W is a monomer that can form a correct (i.e., $A \cdot U$ and $C \cdot G$) or incorrect base pair with a base in a template. The correct base pair is energetically stabler than the incorrect pair. As a rule of thumb, the rate of incorrect addition can be roughly estimated as $\mu \sim 10^{-3}$. Generally, DNA has millions of bases, and therefore heredity of genetic information seems to be almost impossible according to the discussion of error catastrophe in the previous section. How does life achieve accurate replication of information in DNAs?

One of the possible answers is the kinetic proofreading mechanism proposed by Hopfield [92, 93]. Here, we consider the reaction process with a intermediate state in which the catalyst is modified,

$$R/W + E \underset{k^-_{r/w}}{\overset{k^+}{\rightleftarrows}} R/W \cdot E \rightleftarrows R/W \cdot E^* \overset{k}{\to} P_{r/w} + E,$$

$$R/W + E \underset{k^-_{r/w}}{\overset{k^+}{\rightleftarrows}} R/W \cdot E^*.$$

Then, we also suppose that the modification of the catalyst and dissociation of the complex are irreversible processes, i.e.,

$$R/W + E \underset{k^-_{r/w}}{\overset{k^+}{\rightleftarrows}} R/W \cdot E \to R/W \cdot E^* \overset{k}{\to} P_{r/w} + E,$$

$$R/W + E \overset{k^-_{r/w}}{\leftarrow} R/W \cdot E^*.$$

In this case, the proportion of error is square of the original's:

$$\mu \sim (e^{\beta(e_r - e_w)})^2.$$

Similarly, if the number of intermediate states is n, μ is the $n+1$th power.

By such a mechanism, the mutation rate is improved: $\mu \sim 10^{-6}$. Thus, accurate replication of information in DNA becomes possible. It is thought that not only DNA replication but also various reactions and circuits in a living system use such a "design principle" of life. Nonetheless, for the precursor of life, it is not plausible to have such a complicated process with evolutionary consequences.

Generally, in the error correction process, there is a trade-off between production speed and the error rate [94, 95, 96, 97].

1.4 An outline of this thesis

As mentioned above, a living system is a self-replicating system that is composed of many species of biopolymers. How such an autocatalytic system preceding the LUCA had emerged and preserved itself is what we want to study in this thesis.

As reviewed in previous sections, there are many conflicting hypotheses about the origins of life: the RNA world or protein world as hypotheses about the constituent substances of primitive life, and the hypothesis of a warm little pond or hydrothermal-vent hypothesis regarding the place where life has arisen. Based on these hypotheses, there are many attempts to reconstruct the emergence and evolution of primordial life in laboratories.

To serve as the guideline in experimental exploration of self-replicating polymers or to verify the plausibility of such a hypothesis, theoretical analysis of a self-replicating polymer system as a primordial life form is important. Biopolymers that constitute primordial life obey the physical laws of thermodynamics and kinetics. Therefore, it is crucial to discuss such a physical constraint for the elucidation of conditions in which primordial life could have appeared and preserved itself. For example, Dyson analyzed plausibility of the emergence of an autocatalytic system in the phenomenological model. In addition, Eigen demonstrated that the length of heritable information has a limit in the quasispecies model.

In particular, in this thesis, we focus on the kinetic constraints due to the polymerization process of self-replicating biopolymers. Polymerization is the fundamental process in the replication of a biopolymer system, and the synthesis of long and complex polymers inevitably requires several steps of chemical reactions. Under this constraint, what physical property a biopolymer must have to form a self-replicating system that can spontaneously emerge and persist as well as evolve into a protocell that can undergo open-ended evolution? Besides, what condition of the environment is consistent with such an autocatalytic system?

In chapter 2, we will discuss the emergence of an autocatalytic system of a long polymer. As in Dyson's model, the transition from a noncatalytic to catalytic state via stochastic fluctuation will be discussed. we will show that probability of the transition depends on the system size, and there is an optimal one due to the kinetic constraint of polymerization.

In chapter 3, we will discuss the selection of a complex sequence among different kinds of autocatalytic templates in a polymer set. To be a sophisticated autocatalytic system such as prelife, a polymer with a complex monomer sequence is essential for information coding and sophisticated functions. We will see the selection of a specific sequence as a consequence of the kinetic constraint of the reaction pathway to a template sequence. Moreover, it will be revealed that the complexity of the selected sequence depends on the kinetic condition of the environment; i.e., the supply rate of monomers. This result suggests that an optimal nonequilibrium condition in the environment exists for prelife to select and maintain a polymer with complex sequences.

In chapter 4, we will discuss heritability of information in a self-replicating poly-

mer system. As reviewed in the introduction, how to prevent error catastrophe of a long template polymer and to preserve the information of sequences is a big problem for the maintenance of the autocatalytic set. We will examine how the polymerization details of the template affects its error tolerance. We first will see that there is correspondence between the growth law of replicator templates and kinetic parameters of polymerization. In the case where template growth is superexponential, the error tolerance for a mutation increases with the length of the template polymer in contrast to Eigen's quasispecies model in which replication of a template is considered only phenomenologically.

In chapter 5, we will summarize this thesis and discuss its relevance to the origins of life studies.

Chapter 2

Optimal size for emergence of self-replicating polymer system

A biological system consists of a variety of polymers that are synthesized from monomers, by catalysis that exists only for some long polymers. It is important to elucidate the emergence and sustenance of such autocatalytic polymerization. We analyze here the stochastic polymerization reaction dynamics, to investigate the transition time from a state with almost no catalysts to a state with sufficient catalysts. We found an optimal volume that minimizes this transition time, which agrees with the inverse of the catalyst concentration at the unstable fixed point that separates the two states, as is theoretically explained. Relevance to the origins of life is also discussed.

2.1 Introduction

All life systems known so far consist of a wide variety of polymers that catalyze each other and are replicated through catalytic reactions. In cells, for instance, ribosomes whose main component are RNAs that synthesize a variety of protein species, such as polymerases, that catalyze RNA replication [98]. When considering the origins of life, it is therefore necessary to understand the emergence of a primordial polymer system that allows for self-replicating catalytic reactions, in which resource monomers such as amino acids or nucleotides, which are the building blocks of polymers, are supplied [11, 99]. It is also important to understand the timescale of the synthesis of catalytic polymers by polymerizing reactions of the monomers.

In this scenario, a polymer has to be long enough to function as a catalyst. In general, without catalysts, a chemical reaction to synthesize such a long polymer is extremely slow, while polymers, even if they are synthesized, are constantly degraded or diffused out. The synthesis can overcome possible degradation or diffusion only under catalysts (enzyme for protein; ribozyme for RNA) that accelerate the reaction by $10^7 \sim 10^{19}$ [100]. To sustain such a catalytically active state, a certain amount of catalysts is needed, which in turn is only synthesized from catalysts. Hence, the reaction system with autocatalytic polymers is expected to

exhibit bi-stability between the inactive state with almost no catalysts and the active state with abundant catalysts that reproduce themselves. In fact, the importance of the transition from the inactive to active state for the emergence of a primitive replicating system has already been pointed out in the seminal work by Dyson [66], while catalytic reaction networks have also been extensively studied [41, 68, 101, 102, 103, 104]. Here, we study this problem by considering a simple autocatalytic polymerization process, with an aim to obtain the time required for the transition from the inactive to active states.

The existence of bistable states and the transition to a catalytically active state has been discussed recently [69, 80]. In these studies, the rate equation of the concentrations of the monomers and polymers were often adopted. However, at the stage of the emergence of catalytic polymers, the number of molecules may not be so large, and the fluctuations are not negligible. Hence the description by rate equation may not be appropriate. These fluctuations enable the transition across the barrier between the two states. Hence, we adopt a stochastic model with random collisions of molecules to investigate the transition.

The effects of fluctuations due to the smallness in molecule number have attracted considerable attention in chemical reaction dynamics [56, 57, 105, 106, 61, 62, 58]. The change in the steady distribution as well as the relaxation dynamics around the steady state have been studied with respect to the decrease in the system size. However, here, we are interested in the transient time course and statistics for the transition. We study the system-size dependence of the transition time in several models for stochastic polymerization reaction dynamics, in order to find the optimal size that minimizes the transition. We then show that this time is estimated by the inverse of the concentration of the catalytic polymer at the unstable fixed point in the reaction rate equation. We will also explain the origin of this inverse law, and discuss its relevance to the origins of life.

2.2 Autocatalytic polymer model

To introduce our model, we postulate the following properties of the biological polymerization process. (i)Monomers are supplied sufficiently. (ii)Polymerization occurs stepwise from monomers to longer polymers. (iii)Catalytic function can emerge only for polymers with a sufficient length. (iv)Polymerization reaction proceeds extremely slowly without catalysts, but is accelerated drastically with them.

Initially only monomers are supplied, and the polymerization progresses slowly owing to degradation or diffusion. Once sufficient catalysts are synthesized, the catalytic reaction to synthesize them progress constantly. We investigate the emergence and time scale of such autocatalysis by introducing a model consisting of polymers. We denote the polymer of length n as $A(n)$, with $A(1)$ being a supplied monomer. Sequence information under multiple monomer species is disregarded, and only the length is considered. We assume that catalytic capacity appears at $n = L$. For successive polymerization for each ligation reaction, we introduce model (I) with a set of reversible reactions:

$$
\begin{aligned}
A(n-1)+A(1) &\underset{1}{\overset{\alpha}{\leftrightarrows}} A(n) \\
A(n-1)+A(1)+A(L) &\underset{k}{\overset{\alpha k}{\leftrightarrows}} A(n)+A(L) \\
A(n) &\overset{\mu}{\to} \emptyset \qquad (n=2,3,4,\ldots,L)
\end{aligned}
$$

Conversely, for polymerization processes which double the length, we introduce model (II):

$$
\begin{aligned}
A(n)+A(n) &\underset{1}{\overset{\alpha}{\leftrightarrows}} A(2n) \\
A(n)+A(n)+A(L) &\underset{k}{\overset{\alpha k}{\leftrightarrows}} A(2n)+A(L) \\
A(n) &\overset{\mu}{\to} \emptyset \qquad (n=2,4,8,\ldots,L)
\end{aligned}
$$

We first consider the simplest case for model (II), i.e., $L = 4$; the general case will be studied later.

The model consists only of monomers, dimers, and tetramers. Two monomers ligate into a dimer, and dimers ligate into a tetramer. Without losing generality, the spontaneous reaction rate without catalysts is set at unity, while the ligation reaction is accelerated k times by the catalytic tetramers, where $k \gg 1$, accordingly. The backward reaction rate is set to α. We present the case where $\alpha = 1$, while for $\alpha \geq 1$ the result generality holds. Indeed, the synthesis of the catalytic polymer often needs energetic cost, and the forward reaction is slower than the backward one, and then $\alpha \geq 1$ is resulted. Even if $\alpha < 1$, however, as long as the polymer concentration decreases with its length, the discussed results are valid.

These molecules are placed in a container with a (constant) volume. We assume that molecules are well mixed in the container, and we do not consider spatial inhomogeneity. From the external reservoir, monomers are supplied sufficiently fast, such that its number $n_1 (= V)$ is assumed to be constant. Then, our dynamics are represented by the number of dimers $n_2(t)$ and tetramers $n_4(t)$ at time t. The dimers and tetramers are diffused out so that their numbers decrease with the rate μ. Considering that all the reaction processes occur stochastically with the rate following the mass action law for reaction, the dynamics of the probability distribution of n_2 and n_4 is given by

$$
\begin{aligned}
&\partial P(n_2, n_4, t)/\partial t = \\
&\sum_{m=1}^{6} \Big[\pi_m P(n_2 - \nu_{2,m}, n_4 - \nu_{4,m}, t) - \pi_m P(n_2, n_4, t) \Big],
\end{aligned}
\tag{2.1}
$$

where m is the index of the reaction, and $\nu_{2,m}, \nu_{4,m}$ are the increment in the number of dimers and tetramers in the mth reaction (for example, $\nu_{2,1} = 1, \nu_{4,2} =$

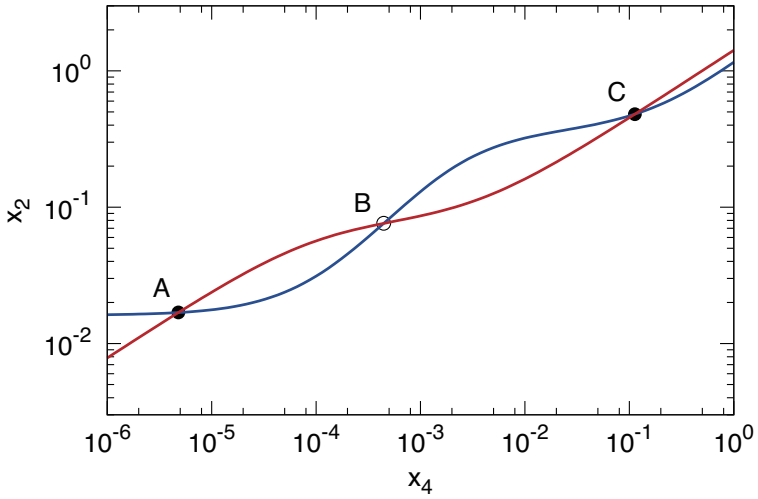

Figure 2.1. Nullclines of the rate equation Eq.2.2 for $k = 10^4$ and $\mu = 30$, plotted in the phase space (x_4, x_2). The blue line represents the nullcline $\dot{x_2} = 0$, and the red line represents the nullcline $\dot{x_4} = 0$. Intersection points of the lines A, B, and C are fixed points of the equation, where A and C (black circle) are stable fixed points and B (white circle) is an unstable one.

$0, \nu_{2,2} = -1, \nu_{4,2} = 0$, and so forth). The first term describes the inflow of the probability from $(n_2 - \nu_{2,m}, n_4 - \nu_{4,m})$ by the reaction, and the second term describes the outflow of probability from (n_2, n_4). Each $\pi_m(\nu_{2,m}, \nu_{4,m})$ is the transition probability of monomers ligation, dimer cleavage, dimers ligation, tetramer cleavage, diffusion of dimer and tetramer, respectively, as given by

$$\begin{aligned}
&\pi_1(+1, 0) = \frac{1}{2}\kappa n_1(n_1 - 1)V^{-1}, \pi_2(-1, 0) = \alpha\kappa n_2 \\
&\pi_3(-2, +1) = \frac{1}{2}\kappa n_2(n_2 - 1)V^{-1} \\
&\pi_4(+2, -1) = \alpha n_4 V^{-1} + \alpha k n_4(n_4 - 1)V^{-1} \\
&\pi_5(-1, 0) = \mu n_2, \pi_6(0, -1) = \mu n_4,
\end{aligned}$$

where κ is the sum of the rate of spontaneous and catalytic reaction as: $\kappa = 1 + k n_4 V^{-1}$.

In the continuum limit with $V, n_i \to \infty$, the change in molecule concentration $(x_1 = 1, x_2 = n_2/V, x_4 = n_4/V)$ [1] is represented by the deterministic rate equation [55], as given by

$$\begin{aligned}
\dot{x}_2 =& (1 + k x_4)(\frac{1}{2}x_1^2 - \alpha x_2 - x_2^2 + \alpha 2 x_4) \quad - \mu x_2 \\
\dot{x}_4 =& (1 + k x_4)(\frac{1}{2}x_2^2 - \alpha x_4) \qquad\qquad\qquad\quad - \mu x_4
\end{aligned} \tag{2.2}$$

[1] Unit of polymer concentration is [number/L]. If we use [mol/L], we need to make a unit conversion by multiplying by the Avogadro constant.

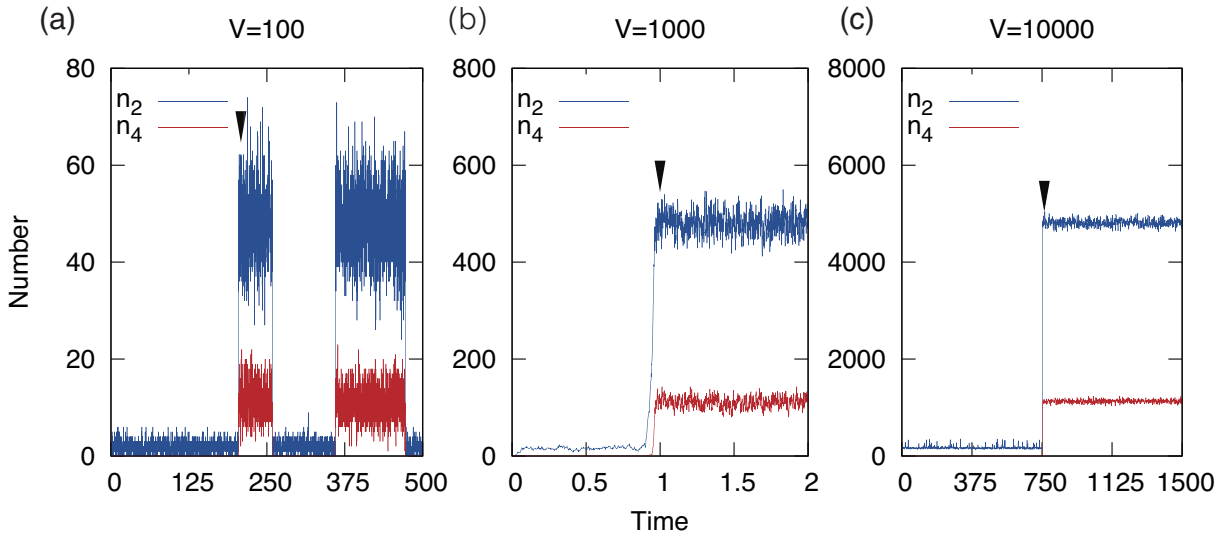

Figure 2.2. Time series of the number of dimers n_2 (blue line) and tetramers n_4 (red line) for $k = 10^4, \mu = 30$ and (a)$V = 10^2$,(b)10^3,(c)10^4. The time for the transition to the active state are $t^* \sim 204$, ~ 1, and ~ 754 for (a),(b), and (c), respectively.

For a certain range of parameters k and μ, this deterministic rate equation has bi-stable fixed points, separated by an unstable fixed point, which are denoted by $A(x_2^a, x_4^a), B(x_2^b, x_4^b)$, and $C(x_2^c, x_4^c), (x_2^a < x_2^b < x_2^c, x_4^a < x_4^b < x_4^c)$ (see Fig.2.1). The fixed point A corresponds to the non-active state with only few dimers and tetramers, while C corresponds to the active state with autocatalytic reactions and sufficient catalysts.

2.3 Optimal volume for the transition and its relation with unstable state

We now study the probabilistic chemical reaction system by the Gillespie algorithm [107]. Examples of the time series of the number of molecules are shown in Fig.2.2, for different values of V, where the initial condition of the number of dimers and tetramers is set to zero. As shown, the transition from the inactive to active state is observed at a certain time, while for small V, there also exists transition back from the active to inactive state.

We now define the transition time to the active state by the time when the concentrations of both dimers and tetramers $n_2/V, n_4/V$ exceed the value given by the active fixed point $C(x_2^c, x_4^c)$ in the rate equation. The volume dependency of

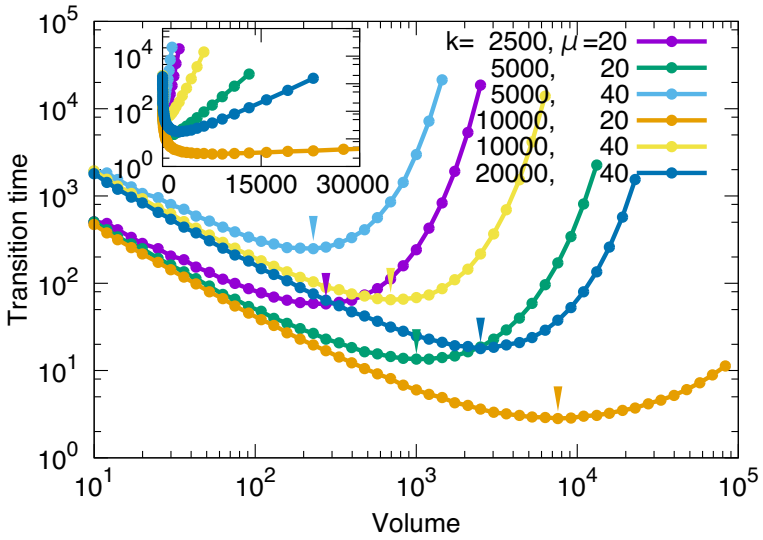

Figure 2.3. The transition time from the inactive to active states, plotted as a function of the volume V. For each parameter and size, the time is computed as the average over 10^4 samples of the master equation. Log–log plot for $(k, \mu) = (2.5 \times 10^3, 20), (5 \times 10^3, 20), (5 \times 10^3, 40), (10^4, 20), (10^4, 40)and(2 \times 10^4, 40)$, with different colors and symbols. The optimal V_{opt} that gives the minimum is indicated by the thick arrow. The inset shows the corresponding semi-log plot.

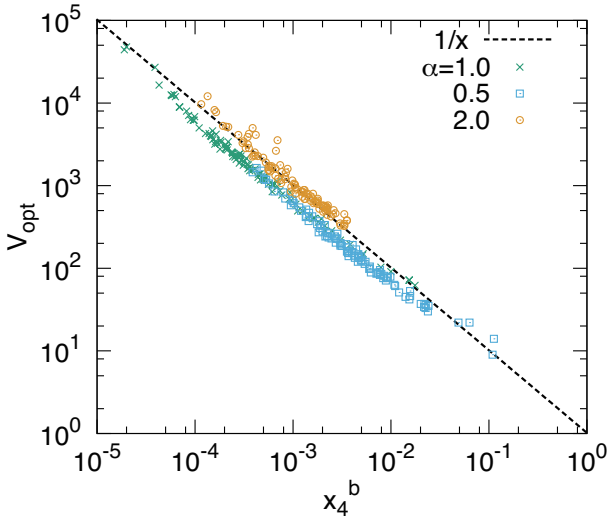

Figure 2.4. The relationship between V_{opt} and x_4^b. The abscissa give the tetramer concentration at the unstable fixed point x_4^b, and the ordinate shows the optimal volume V_{opt} as estimated in Fig.2.3. Each point is a result from different parameter values randomly selected from $k \in (0, 5 \times 10^4), \mu \in (0, 100)$ where the rate equation exhibits bi-stability.

the transition time is plotted in Fig.2.3. As can be seen in the figure there is a minimum transition time at a certain optimal volume V_{opt}. By decreasing V below V_{opt}, the transition time increases asymptotically as V^{-1}, while for $V > V_{opt}$, it increases exponentially with the volume.

We computed V_{opt} by varying the parameters k and μ. Interestingly, as shown in Fig.2.4, these optimal volumes, when plotted as a function of x_4^b, i.e., the tetramer concentration at the unstable fixed point in rate equation, are fitted on a single curve, given by $V_{opt} \sim 1/{x_4^b}^{\,2}$, for all the parameter values. (If backward reaction rate α is close to zero, the polymer concentration could be higher than the monomer at the unstable fixed point, and in this case, the monomer number that satisfies $1/x_4^b$ is less than unity for small V, and the optimal volume cannot exist. Apart from this unrealistic case this relationship of the optimal volume generally holds.)

We now explain the origin of this optimal volume and its relationship with x_4^b. First of all, by using the standard formalism [55] master equation is approximated by the Fokker–Planck equation (FP) for large volume. It is obtained by the Kramers–Moyal expansion or the chemical Langevin equation, as the expansion by $1/V$. For our model ($\alpha = 1, x_1 = 1$), the FP equation is straightforwardly obtained as

$$\begin{aligned}
\partial P(x_2, x_4, t)/\partial t = &- \sum_{i=2,4} \partial_{x_i}(A_i(x)P(x,t)) \\
&+ \frac{1}{2V} \sum_{i,j=2,4} \partial_{x_i}\partial_{x_j}(B_{i,j}(x)P(x,t)) \\
A = &\begin{pmatrix} (1+kx_4)(\frac{1}{2} - x_2 - x_2^2 + 2x_4) - \mu x_2 \\ (1+kx_4)(\frac{1}{2}x_2^2 - x_4) - \mu x_4 \end{pmatrix} \\
B = &\begin{pmatrix} (1+kx_4)(\frac{1}{2} + x_2) + \mu x_2 & 4(1+kx_4)(\frac{1}{2}x_2^2 + x_4) \\ (1+kx_4)(\frac{1}{2}x_2^2 + x_4) & \mu x_4 \end{pmatrix}
\end{aligned} \tag{2.3}$$

We computed the volume dependency of the transition time from the initial value $x_2 = x_4 = 0$ to $x_2 > x_2^c, x_4 > x_4^c$ by using the FP equation, as is plotted in Fig.2.5. The transition time increases monotonically with V, and is fitted well with $\exp(const. \times V)$. This is expected from the Kramers formula [54] in which the probability to jump from one stationary state to another is given by $\exp(-\epsilon U)$, where ϵ is the noise strength and U is the potential barrier to go to the new stationary state. Here the noise in the FP equation is proportional to $\epsilon = 1/V$, and the transition to the active state which has to go across the unstable fixed point is given by the jump over the potential barrier U_b. Hence the transition time increases as $\exp(U_b V)$ with the increase in V. In fact, the transition time computed by the master equation asymptotically agrees with this estimate for large V.

(Indeed, for a one-variable Fokker–Planck equation

$$\partial_t P(x,t) = -\partial_x\big(A(x)P(x,t)\big) + \frac{1}{2}\partial_x^2\big(B(x)P(x,t)\big), \tag{2.4}$$

[2]With the unit of concentration [mol/L], this expressions is written as $V_{opt} \sim 1/(x_4^b N_A)$.

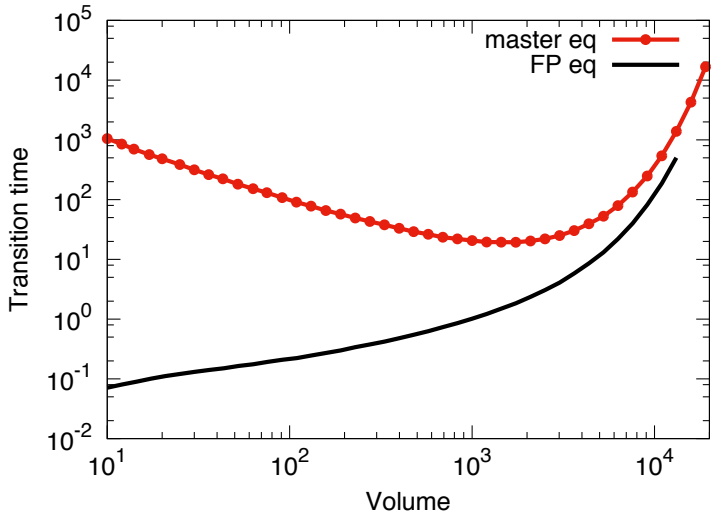

Figure 2.5. The transition time estimated by the FP equation Eq.2.3 for $k = 10^4$,$\mu = 30$ (black line); the transition time directly estimated by the master equation is also shown for comparison (red line).

the corresponding potential is given by $U(x) = \int^x A(x')/B(x')dx'$. In the present two-variable case, such potential does not generally exist [108], but the temporal evolution to cross the saddle-point B is restricted around its unstable manifold that is located between the two nullclines (see Fig.2.1). Thus the motion is effectively represented by a one-dimensional motion, and the effective potential exists for the present Fokker–Planck equation.)

As this dependence is monotonic in V, the FP equation can explain neither the existence of V_{opt} nor the increase in the transition time for $V < V_{opt}$. For the explanation the discreteness of the molecule number $0, 1, 2, \ldots$ is very important. We now explain the relationship $V_{opt} \sim 1/x_4^b$ along this line.

For the transition to occur, the concentration x_4 has to exceed the value of the unstable fixed point x_4^b, while the catalytic reaction needs at least a single tetramer. For small $V < 1/x_4^b$, the average tetramer number is decreased below unity. The tetramer number is therefore often zero, and the probability of having a single molecule decreases with the decrease in V. Then the probability of crossing x_4^b decreases with V, if $Vx_4^b < 1$. Thus, the transition time increases as V decreases below $1/x_4^b$, which gives the optimal volume for the transition time.

We have confirmed that the relationship between the optimal volume and the unstable fixed point is universal in an autocatalytic polymerization process, as given by models (I) and (II). Indeed, most autocatalytic polymerization process can be essentially constructed by combination of models (I) or (II). Here, the master equations for probabilities are given in the same way as in Eq.2.1, and the rate equation

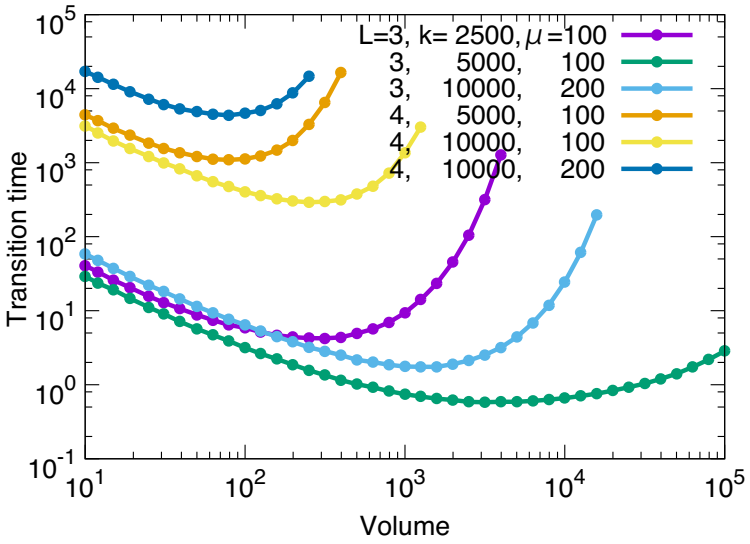

Figure 2.6. The same plot for model (I) with $(L, k, \mu) = (3, 2.5 \times 10^3, 100), (3, 5.0 \times 10^3, 100), (3, 1.0 \times 10^4, 200), (4, 5.0 \times 10^3, 100), (4, 1.0 \times 10^4, 100), (4, 1.0 \times 10^4, 200)$. Computed in the same manner as Fig.2.3

in the infinite volume limit for models (I) and (II) is given by:

$$\begin{aligned} \dot{x}_n &= (1 + kx_L)(x_{n-1}x_1 - \alpha x_n - x_n x_1 + \alpha x_{n+1}) - \mu x_n \\ &\qquad (n = 2, 3, 4, \ldots, L-1) \\ \dot{x}_L &= (1 + kx_L)(x_{L-1}x_1 - \alpha x_L) - \mu x_L, \end{aligned} \tag{2.5}$$

for model (I) and

$$\begin{aligned} \dot{x}_n &= (1 + kx_L)(\frac{1}{2}x_{n/2}^2 - \alpha x_n - x_n^2 + 2x_{2n}) - \mu x_n \\ &\qquad (n = 2, 4, 8, \ldots, L/2) \\ \dot{x}_L &= (1 + kx_L)(\frac{1}{2}x_{L/2}^2 - \alpha x_L) - \mu x_L. \end{aligned} \tag{2.6}$$

for model (II).

Again, the dynamical systems of the model equations (2.5) and (2.6) have two stable fixed points $A(x_i^a), C(x_i^c)$ and one unstable fixed point$B(x_i^b)$, for a certain parameter region. Fixed points in model (II) are the root of the self-consistent equation:

$$x_L = \left(\frac{1 + kx_L}{2\alpha(1 + kx_L) + 2\mu}\right)^{L-1} x_1^L.$$

Using direct Gillespie simulations of the model, we computed the transition time between the state with $x_L \sim 0$ to the active state with $x_L \sim x_L^c$, which has a minimum at a certain volume V_{opt} (see Fig.2.6). We have plotted V_{opt} for a variety of parameters and maximal lengths L, again as a function of x_L^b in Fig.2.7. We have

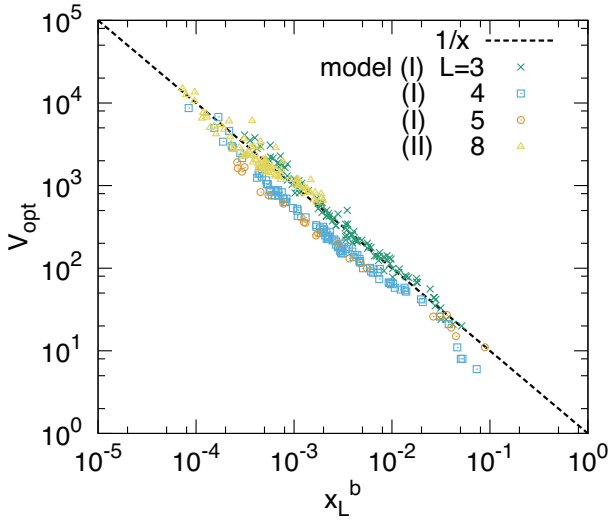

Figure 2.7. The relationship between V_{opt} and x_L^b for model (I) with $L = 3$ (green), 4 (blue) and 5 (orange) and for model (II) with $L = 8$ (yellow). The data are computed in the same way as for Fig.2.4. Parameter values are randomly selected from $k \in (0, 10^4), \mu \in (0, 100)$.

thus confirmed that the relationship between the optimal volume and the catalyst concentration at the unstable fixed point, $V_{opt} \sim 1/x_L^b$ is universal.

Last, we have also studied a model with plural monomer species and sequence dependent catalytic reaction. In this case again, we have confirmed the existence of optimal size for the transition to the active catalytic state, where specific sequences are selected.

2.4 Summary and Discussion

To sum up, we have unveiled the explicit condition for autocatalytic polymers to emerge. As spontaneous synthesis of polymers is much slower than the diffusion or degradation, longer polymers are not maintained only by that. In contrast to this inactive state, catalysts are continuously replicated with the autocatalytic reaction, in the active state. Fluctuations in the number of molecules induce the transition from the former to the latter state while at least a single catalyst is needed for it. This tradeoff leads to the optimal volume to minimize the transition time, generally given by the inverse of the catalyst concentration at the unstable fixed point of the deterministic rate equation. Below this optimal volume, the time to assemble a single catalytic polymer increases in proportion to V^{-1}, while beyond it, more catalytic polymers are needed for the transition, and the time to achieve such fluctuation increases exponentially with V.

Discreteness (0,1,..) in the molecule number essentially matters for the determination of the optimal volume, in contrast to several studies that adopt the FP equation by the system-size expansion [55, 109] to obtain the change in distribution

[61, 58] and the transition process [60]. In fact, the FP equation approach can account for the exponential increase in the transition time with the system size but not the optimal volume. It is also interesting to note that despite the importance in the discreteness in the number, the optimal volume is estimated by the unstable fixed point of the rate equation that itself is obtained in the infinite size limit.

Our result applies generally to an autocatalytic replication process, where inactive and active states are bistable. The inverse relationship between the optimal volume and the catalyst concentration at the unstable fixed point that separates the two stable states is general even though the unstable-point concentration depends on several reaction parameters, and the details of the polymerization process. In models (I) and (II), the concentration decreases exponentially with the (minimum) length of the catalyst polymer, such that the optimal volume as well as the transition time increases.

Although the emergence of life will require several steps, autocatalytic polymerization is essential as one of them. The present result suggests that the volume of the reaction space matters for it. A limited space with an appropriate size is preferable, which would be first provided by a porous medium such as a mineral surface, and later by a vesicle composed by the lipid bilayer. For example, by considering the model (II) with $k = 10^8$[L/mol], $\mu = 1$ (with the unit of spontaneous reaction rate), $\alpha = 1$[L/mol], $x_1 = 1$[mol/L] and $L = 16$ the unstable fixed point value of the catalyst in the rate equation becomes $x_L^b = 8.4 \times 10^{-9}$[mol/L]. Therefore, $V_{opt} \sim 1/(x_L^b N_A)$ is estimated by 2×10^{-16}[L], which is as small value as a bacterial cell, while these parameter values are not unique and reliable estimates are difficult since the chemical parameters for the primitive catalytic polymerization are not currently available. This estimate, however, will be useful to design protocells with catalytic polymers within a vesicle, which has been extremely investigated in synthetic biology [110, 111]. Our optimal size will provide a guide to choose an appropriate vesicle size.

Chapter 3

Kinetic Selection of Template Polymer with Complex Sequences

Emergence and maintenance of polymers with complex sequences pose a major question in the study of the origins of life. To answer this, we studied 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 revealed 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 lower flow rate. We discuss the relevance of this kinetic sequence selection through nonequilibrium flow to the origin of complex polymers.

3.1 Introduction

Schrödinger, in his celebrated monograph "What is Life" [112], recognized that the essence of 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 origins of life or self-replication, catalytic polymers are required, as has also been investigated in recent experiments [50, 34, 113, 35, 114]. In their study, polymers with complex sequences had to be synthesized and preserved, to encode a large amount of information and for catalytic functions [115, 116]. 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 auto-catalytic in nature, as has been extensively investigated [117, 68, 118, 104, 69, 119, 70, 120], including theoretical models that explicitly consider template-like self-replicating polymerization

[72, 121, 74, 77, 78, 76, 116, 122, 79] and some synthesis experiments [114, 123, 48].

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. 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 origins of life and by a chemostat condition in a laboratory experiment [42]. 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 [72, 121, 74, 77, 78, 76, 116, 122, 79] does not show any dependence on external conditions. Ref [120] 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 *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.

3.2 Template directed polymer model

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 (Fig. 3.1). Thus, all polymer species are described by the binary integer s of length l (for example, 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.

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., $1+00 \to 100$ and $00+1 \to 001$). 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 $111 + 1 \to 1111$ is catalyzed by the template containing

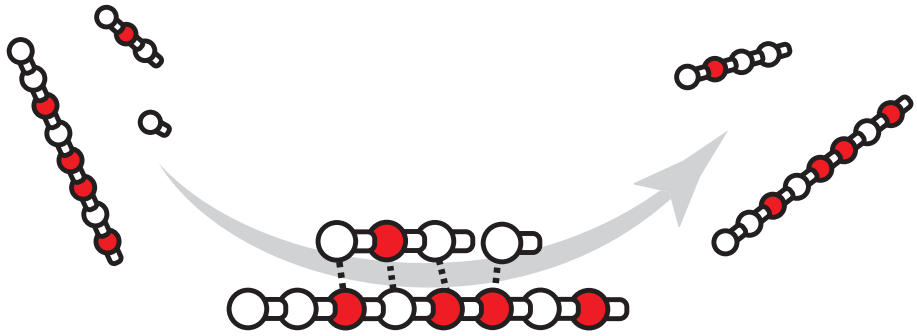

Figure 3.1. The catalytic ligation reaction of a polymer with monomers. Circles with different colors represent different types of monomers. The arrow indicates the direction in which the reaction proceeds. In the reaction highlighted by the arrow, a polymer 010 and a monomer 0 ligate into the product polymer 0100. The reaction is catalyzed by a template polymer 00101101, which contains the complementary sequence of the product 1011 as a subsequence.

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 $A_{l,s}$:

$$\begin{aligned} \phi &\xrightarrow{\frac{1}{2}fV} m, \quad & A_{l,s} + m &\xrightarrow{\kappa_{l+1,sm}} A_{l+1,sm}, \\ A_{l,s} &\xrightarrow{d} \phi, \quad & m + A_{l,s} &\xrightarrow{\kappa_{l+1,ms}} A_{l+1,ms}, \end{aligned}$$

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. (For $l = 1$ the top and the bottom reactions are considered only once. The reverse ligation reaction is neglected here, but its inclusion does not alter the following results qualitatively, as long as its rate is small.) The polymerization can occur spontaneously at a rate, ϵ, 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.

$$\begin{aligned}
\dot{x}_{1,m} =& \frac{1}{2}f - x_{1,m}\sum_{l',s'}(\kappa_{l'+1,s'm} + \kappa_{l'+1,ms'})x_{l',s'} - dx_{1,m}, \\
\dot{x}_{l,s} =& \kappa_{l,s}x_{1,m}(x^{(1)}_{l-1,s} + x^{(2)}_{l-1,s}) \\
& - x_{l,s}\sum_{m'} x_{1,m'}(\kappa_{l+1,sm'} + \kappa_{l+1,m's}) - dx_{l,s}, \\
\dot{x}_{L,s} =& \kappa_{L,s}x_{1,m}(x^{(1)}_{L-1,s} + x^{(2)}_{L-1,s}) - dx_{L,s}.
\end{aligned} \tag{3.1}$$

Here, $x^{(1)}_{l-1,s}$ and $x^{(2)}_{l-1,s}$ are the concentrations of $(l-1)$-mer, which is a subsequence of $A_{l,s}$ on both the sides (For $l = 2$, eq. 1 should be read with $x^{(1)}_{1,s} = x_{1,m}$ and $x^{(2)}_{1,s} = 0$).) and $\kappa_{l,s} = \frac{1}{2}(\epsilon + \sum_{(l',s')\in\mathcal{T}_{l+1,s}} n_{l',s'}/V)$, $\mathcal{T}_{l,s}$ 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 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.

3.3 Results of stochastic simulation of chemical reaction system

If the system size is finite, fluctuations around the rate Eq. 3.1 induce switching among the attractors, and specific attractor(s) are selected dominantly. We investigated this selection by numerically solving the original stochastic reaction model [107]. In Fig. 3.2, 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 chapter.

Next, the temporal average of the concentration of each template sequence was computed to examine its dependence on the flow rate, f (Fig. 3.3), 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 depends 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 Fig. 3.4(a), the result showed a stepwise switch of the dominant sequence species; the dominant sequence changes as 010101 $\rightarrow$ 011001 $\rightarrow$ 001001 $\rightarrow$ 000100 with a decrease in f.

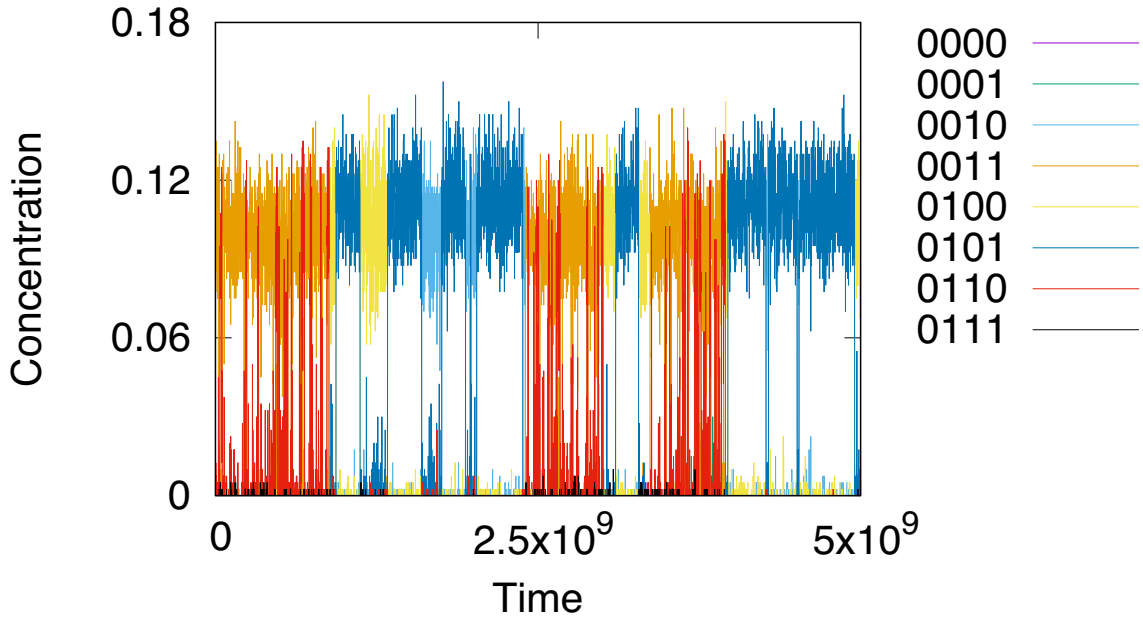

Figure 3.2. Time series of concentrations of all template species. $L = 4, V = 400, f/d = 1, \epsilon = 10^{-4}$, $f = 7.0 \times 10^{-5}$.

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 complimentary ones. For example, 0101 includes only 01 and 10 dimers, while 0010 includes 00, 01, 10, and 11, and is thus more complex. In general, let $\mathcal{C}_l(2k)$ 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 $\mathcal{C}_2(2)$ because it includes two dimer subsequences 01 and 10, and 0110 belongs to $\mathcal{C}_2(4)$ 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 $\mathcal{C}_l(2k)$. It is defined by $x_l(2k) = \sum_{s \in \mathcal{C}_l(2k)} x_{L,s}$. In Fig. 3.4(b), $x_l(2k)$ is plotted as a function of f. For $L = 4$, the concentration of the dominant sequence changes from $x_2(4)$ to $x_2(2)$, whereas for $L = 6$, it changes 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.

(Note that as the flow rate reaches zero, the system approaches thermal equilibrium, where the frequency of each template sequence tends to be uniform if the backward reaction is also taken into consideration. The number of sequences that belong to $\mathcal{C}_3(6)$ is the largest among all the hexamers, and therefore, in Fig. 3.4(b), the frequency $x_3(6)$ increases as the flow rate approaches zero.)

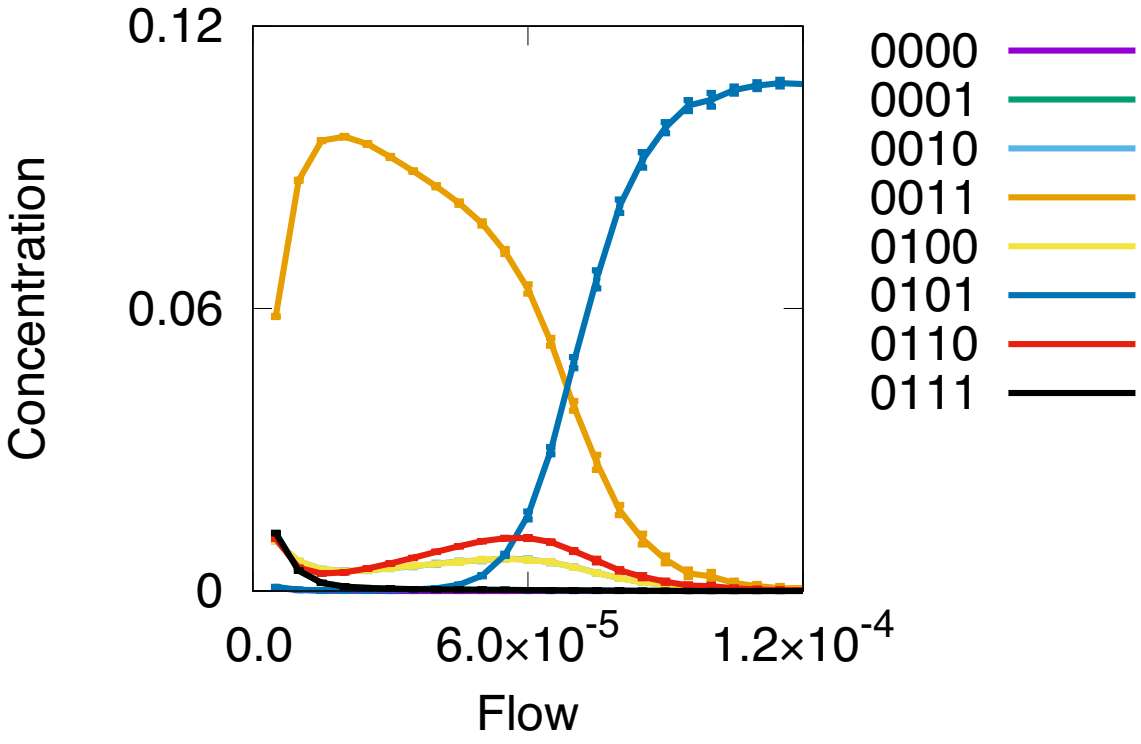

Figure 3.3. Average concentration of each template sequence in stationary state. The horizontal axis is the flow rate in chemostat $f(= d)$. The time intervals for average are varied from 10^{13} to 10^{14} depending on f. A sequence and its left-right reverse take the same average concentration over long time, and the corresponding two lines are overlapped.

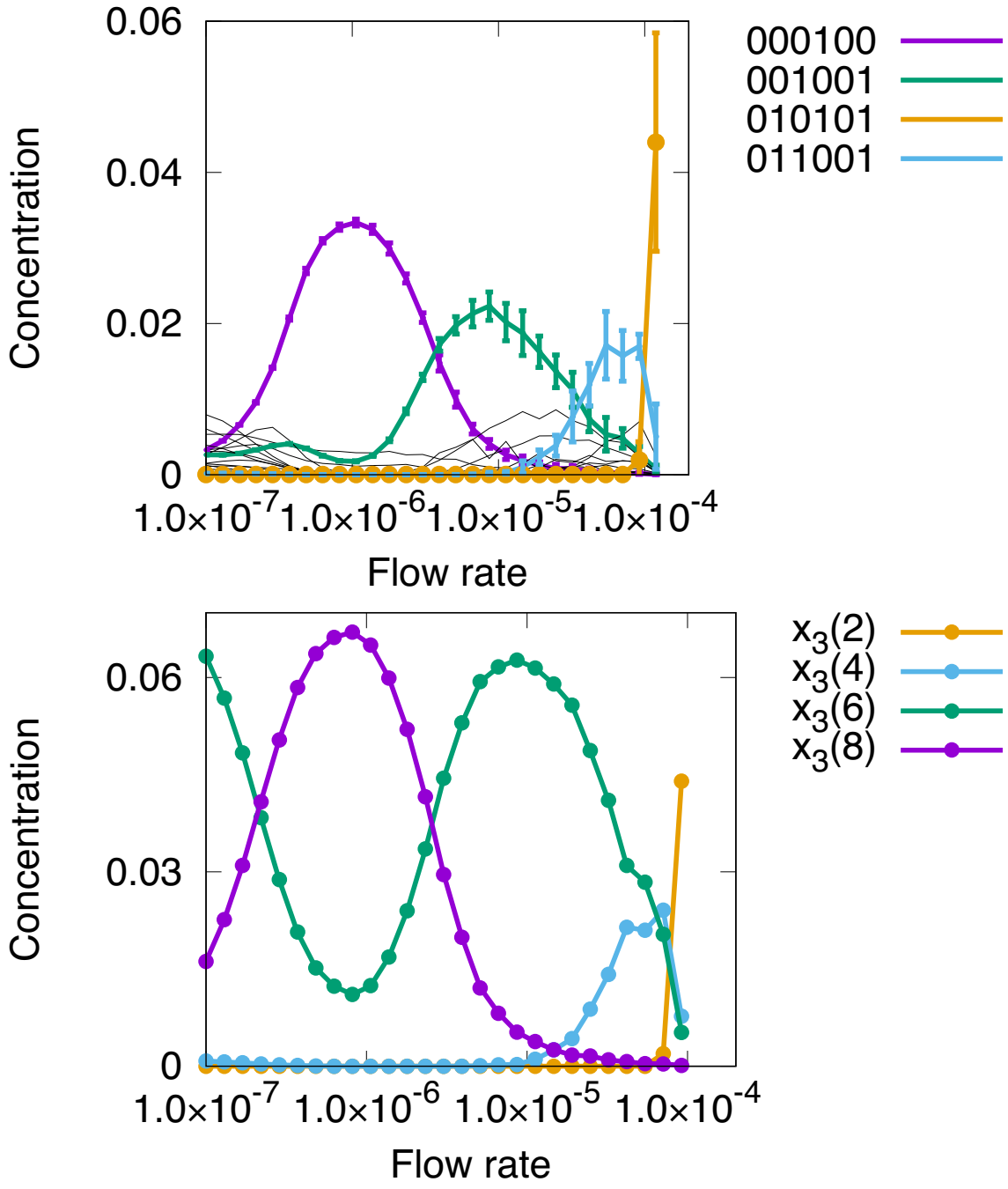

Figure 3.4. (a)Flow rate dependence of the average concentration of hexamer templates in stationary state. The horizontal axis is the flow rate in chemostat $f(=d)$. We set $L = 6, V = 500, \epsilon = 5.0 \times 10^{-6}$. The most dominant sequence is switched stepwise with the decrease in flow rate, as 010101 $\rightarrow$ 011001 $\rightarrow$ 001001 $\rightarrow$ 001000, which corresponds to $\mathcal{C}_3(2)$, $\mathcal{C}_3(4)$, $\mathcal{C}_3(6)$, and $\mathcal{C}_3(8)$, respectively. For the flow rate 10^{-7}, the distribution approaches that of thermal equilibrium.

3.4 Analysis by dynamical systems and reaction pathway to synthesize templates

To understand this change in the dominant sequence, we note that the residence time for each attractor, under noise due to the finiteness in 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 $\mathcal{C}_2(4)$ to that with $\mathcal{C}_2(2)$. 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. 3.5(a), where the change from $\mathcal{C}_2(2)$ with 0101 to $\mathcal{C}_2(4)$ 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 0101 and 1010 pair (denoted as B) and that with 0010 and 1101 pair (as A) by focusing only on these two pairs of templates, while neglecting other template polymers. (This approximation is valid when the rate of spontaneous ligation is much smaller than the time scale of the transition dynamics.) 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:

$$\dot{x}_A = r_A(x_A, x_B)x_A - dx_A, \tag{3.2}$$

$$\dot{x}_B = r_B(x_A, x_B)x_B - dx_B. \tag{3.3}$$

Here, r_A (r_B) is the synthesis rate of the corresponding template for each type A (B) for a given x_A, x_B, respectively, whose form is obtained from Eq. 3.6, as shown in the Appendix. This rate equation has two fixed-point attractors, $(x_A = x_A^*, x_B = 0)$ and $(x_A = 0, x_B = x_B^*)$. (To analyze the stability of each attractor, the (tiny) spontaneous ligation term is not necessary, and it is neglected.) The vector field for dx_A/dt, dx_B/dt is thus obtained as in Fig. 3.6. Here, the attractor close to the unstable fixed point is each 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 Eq. 3.4, 3.5 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 Appendix). 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 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 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 Appendix for analytic derivation of this dependence).

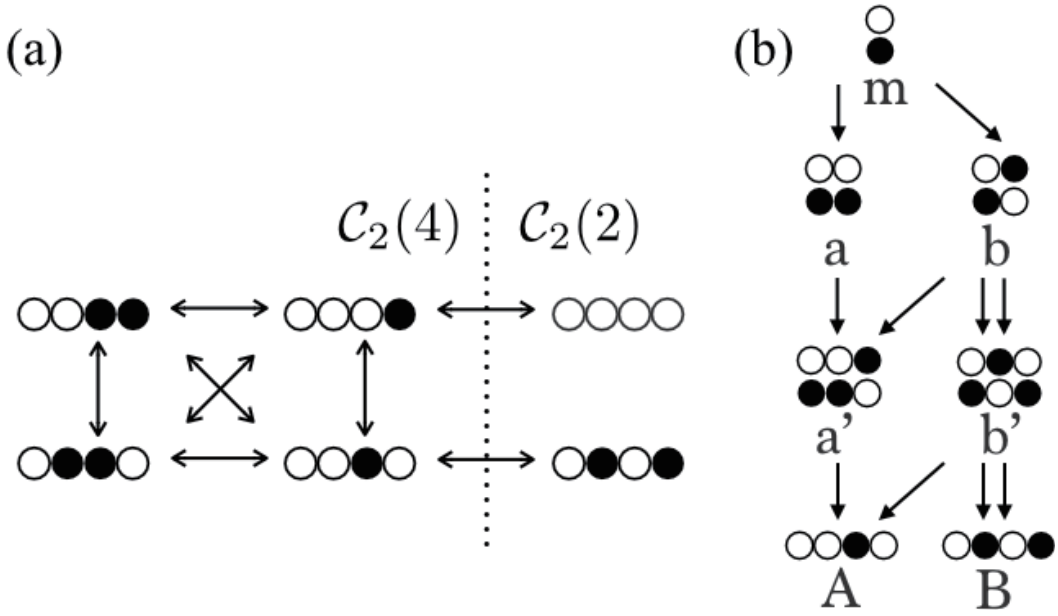

Figure 3.5. (a) Transition network between attractors. Double head indicates transition between them and each sequence represents the one of a complementary pair that is dominant at each attractor. (b) Schematic diagram representing the reaction pathway from monomers to each template A and B. White and black circles represent 0 and 1 monomers, respectively. Arrows indicate 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.

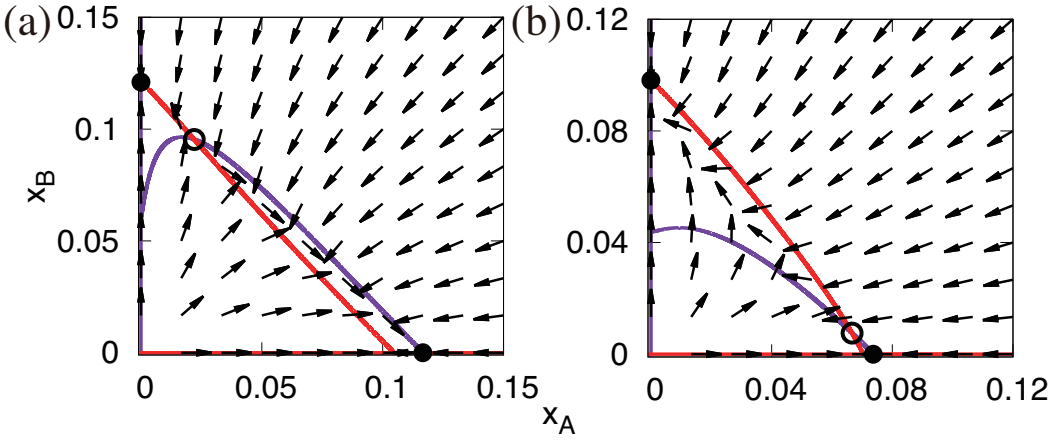

Figure 3.6. Vector field of x_A and x_B. Blue line is the nullcline of $\dot{x}_A = 0$ and red is the nullcline of $\dot{x}_B = 0$. The nullclines overlap x_A line and x_B line partially. In both the figures, there are two stable fixed points (•) and one unstable fixed point (◦). Two stable fixed points are A-abundant and B-abundant attractors, respectively. As the value of f increases from $f(= d) = 1.0\times10^{-5}$(a) to 5.0×10^{-4} (b), the unstable fixed point moves toward the A-abundant attractor.

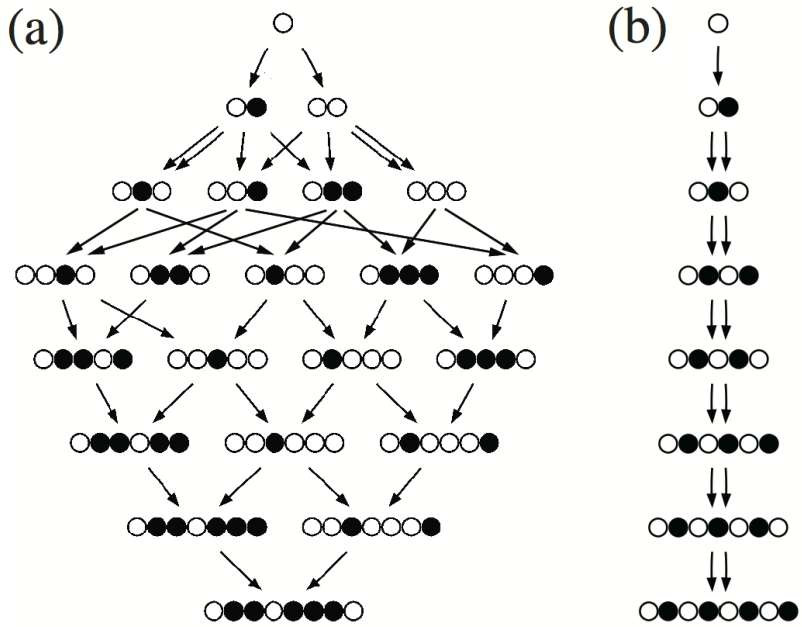

Figure 3.7. Reaction pathway to synthesize 8-mer templates from monomers for (a) a complex sequence (01101110 $\in \mathcal{C}_4(10)$) and (b) a simple sequence (01010101 $\in \mathcal{C}_4(2)$). Black and white circles represent monomers 1 and 0, respectively. Each arrow indicates ligation reaction between a shorter polymer and a monomer, which are catalyzed by product templates.

The above argument holds for all the template sequences belonging to $\mathcal{C}_2(2)$ and $\mathcal{C}_2(4)$. 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 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 complex sequence is dominant for a low f. In Fig. 3.7, 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 direct simulation is plotted in Fig. 3.8 as a function of f.

Since direct simulation of cases with larger L is time-consuming, we studied a reduced model, where only the 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 Appendix for details). A complex sequence with a large number of subsequences was selected as the flow rate decreased.

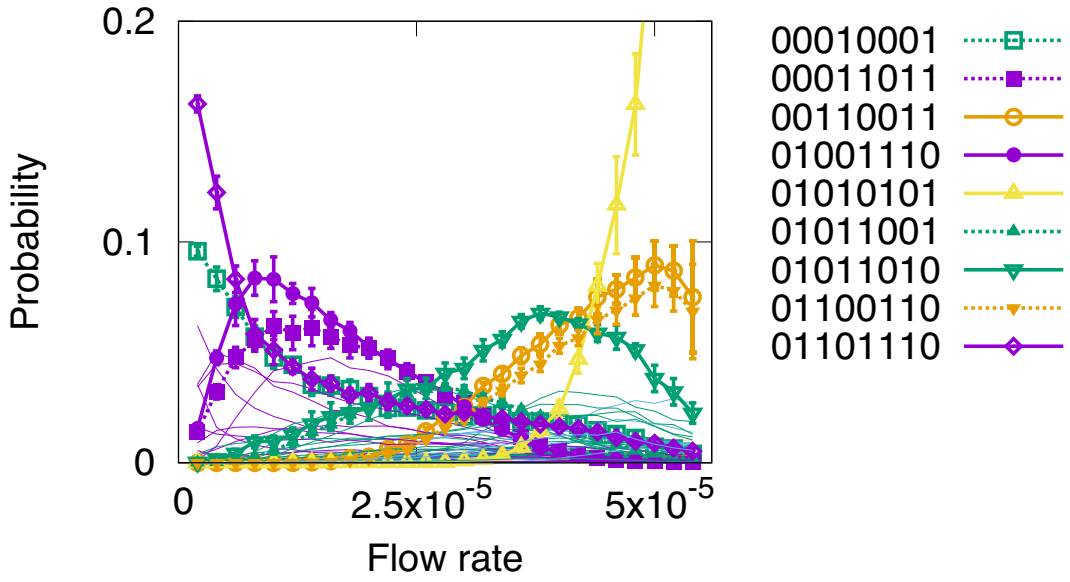

Figure 3.8. Residence probability of attractor at which a pair of sequences dominant. We set the length of the template L as 8 and $\epsilon = 5.0 \times 10^{-6}$, $V = 500$. The major sequences are plotted as thick lines. The probability is calculated as time average of $10^{14} \sim 10^{15}$ simulation time. One of a reverse pair (e.g. 01101110 and 01110110) is omitted here. The complexity of the dominant sequence decreases with decreasing flow rate: The dominant sequence changes from 01010101 ($\mathcal{C}_4(2)$), 00110011 ($\mathcal{C}_4(4)$), 01011010 ($\mathcal{C}_4(6)$), to 01001110 or 01101110 ($\mathcal{C}_4(10)$) with a decrease in flow rate.

3.5 Generality of the results

We have confirmed the generality of the present result, i.e., successive increase in the complexity of dominant sequence with the decrease in the flow rate, for the following cases: (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 [124]).

(i) Variation in catalytic activity

We tested the case in which the catalytic activity of each template L for the sequence s is not identical, given by $k_{L,s}$, so that $\kappa_{l,s}$ in the main text is replaced by $\frac{1}{2}(\epsilon + \sum_{(L,s')\in\mathcal{T}_{l+1,s}} k_{L,s'} n_{l',s'}/V)$. As an example, we studied the tetramer as the template, by changing the catalytic activity of the template 0101 (and 1010), which is the dominant sequence for fast flow, so that the activity for the sequence is given by k while that for other sequences is set at unity.

In Fig. 3.9, we plotted the fraction of each sequence, as shown in Fig. 3.3 The dominant templates change depending on the flow rate, whereas the transition flow rate and the frequency of dominant species changes so that the fraction of sequence 0101 is increased (decreased) with the increase in k. If k is too large (say ~ 2) (too small (say ~ 0.5)), 0101 remains (is not) the dominant species, respectively, against the change in the flow rate. The competition between the difference in the catalytic activities and sequence diversity can be compared to the competition between the energy and entropy in equilibrium thermodynamics.

(ii)Change in the ligation speed

To examine the generality of the results in the main text, i.e., the increase in sequence complexity with the decrease in flow rate, we studied an alternative model in which the rules for catalytic strength were changed from that in the original model.

In the alternative chemical reaction system, the ligation reaction rate of l-mer polymer with sequence s ($A_{l,s}$) $\kappa_{l,s}$ is changed as follows.

$$\kappa_{l,s} = \frac{1}{2}(\epsilon + \sum_{(L,s')\in\mathcal{T}_{l+1,s}} m_{l,s,L,s'} n_{L,s'}/V),$$

where $m_{l,s,L,s'}$ is the number of l-mers with sequence s, which is contained in the template with sequence s'. Under this rule, for example, the tetramer sequence

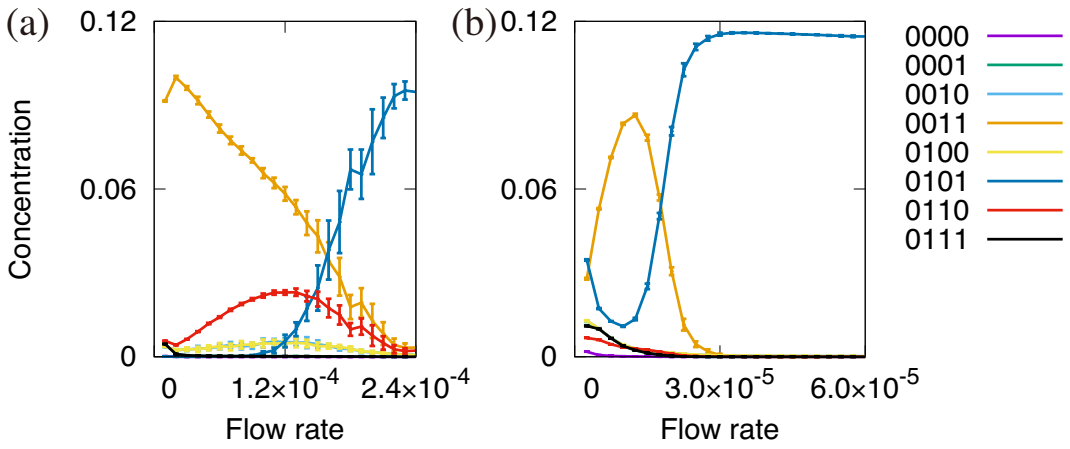

Figure 3.9. Average concentration of each template sequence in the stationary state. The horizontal axis is the flow rate in chemostat $f(=d)$. $L = 4, V = 400, f/d = 1.\epsilon = 10^{-4}$. k for the sequence 0101 (or 1010) is 0.9(a), 1.4(b) while k_s for other sequences is kept at unity.

0101 has two dimers 01 as subsequences and catalyzes the reaction $0+1 \to 01$ three times faster and $1+0 \to 10$ two times faster as compared to the catalytic strength in the original rule.

We calculated the flow rate dependence of the average concentrations of templates in the stationary state (Fig. 3.10). Although the dominant sequence at each flow rate differs from that of the original rule (e.g. in the high flow rate case, not only 0101 but also 0000 is selected), a stepwise increase in the complexity of the dominant sequence with a decrease in flow rate is observed, as in the original model.

(iii)Ligation between polymers beyond the monomer

We also simulated another model in which there is not only ligation between a monomer and a polymer but also ligation between two polymers longer than the monomers (e.g. $01+01 \to 0101$, $01+0101 \to 010101$), so that the ligation reaction in the original model is replaced by

$$A_{l,s} + A_{l',s'} \xrightarrow{\kappa_{l+l',ss'}} A_{l+l',ss'}.$$

In this case also, the flow rate dependence of the complexity of the dominant sequence is observed, as in Fig. 3.11.

(iv) Difference between concentration of monomers 0 and 1

In the main text, we assumed that the ratio between monomer species 0 and 1 was equal. Here, we also calculated the case in which the supply rates of the two monomers are different, by varying the flow rate of monomers 0 given by f_0, and that of monomers 1 given by f_1, while keeping $f_0+f_1 = 1$ fixed. The flow rate dependence of the averaged concentration of template species is shown in Fig. 3.12. If the bias

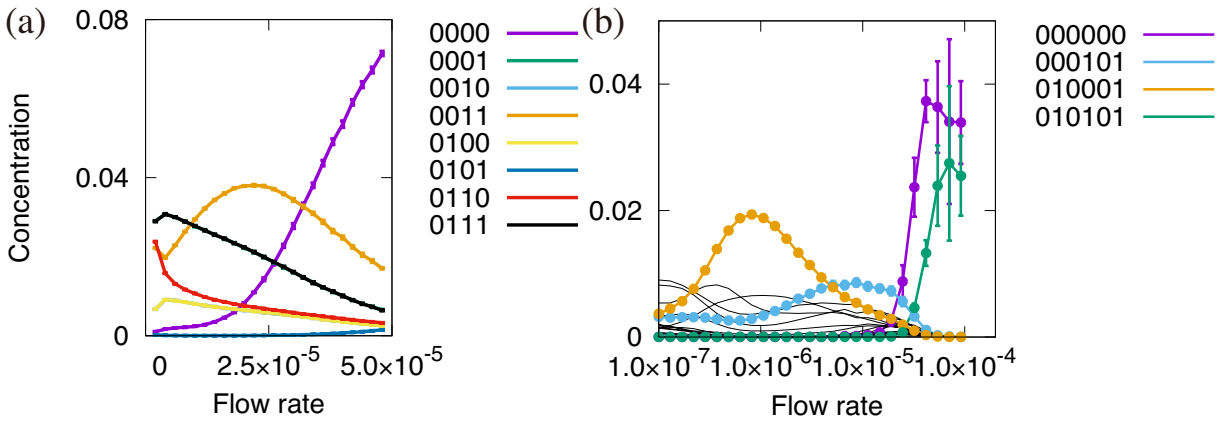

Figure 3.10. (a)Flow rate dependence of average concentrations of templates in the stationary state. The horizontal axis is the flow rate in chemostat $f(=d)$. We set $L = 4, V = 200, \epsilon = 1.0 \times 10^{-5}$ (a) and $L = 6, V = 250, \epsilon = 5.0 \times 10^{-6}$(b). The most dominant sequence is switched stepwise with the decrease in flow rate as 000000 or 010101 $\rightarrow$ 000101 $\rightarrow$ 010001, corresponding to $\mathcal{C}_3(2)$, $\mathcal{C}_3(6)$, and $\mathcal{C}_3(8)$, respectively.

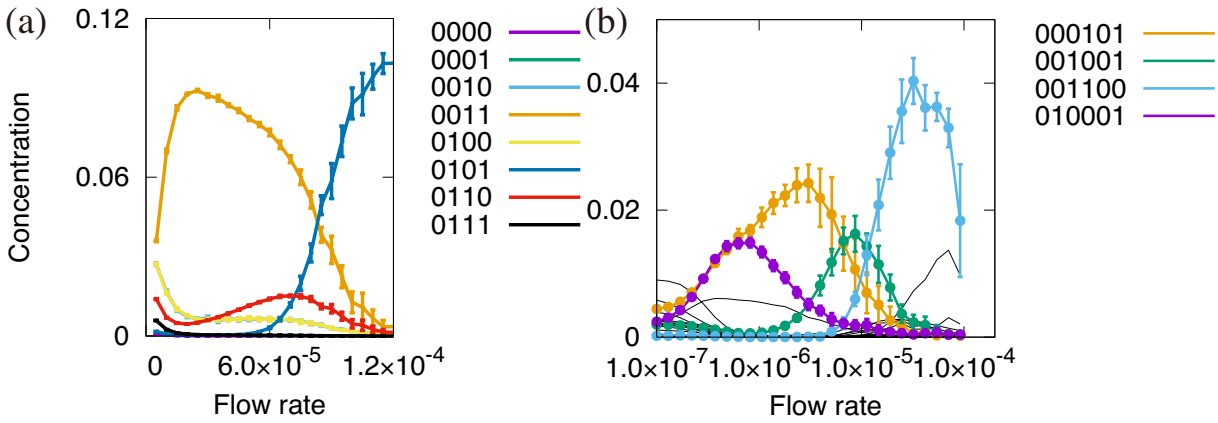

Figure 3.11. Flow rate dependence of average concentrations of templates in the stationary state. The horizontal axis is the flow rate in chemostat $f(=d)$. We set $L = 4, V = 400, \epsilon = 1.0 \times 10^{-4}$ (a) and $L = 6, V = 250, \epsilon = 1.0 \times 10^{-6}$. The most dominant sequence is switched stepwise with the decrease in flow rate as 001100 $\rightarrow$ 001001 $\rightarrow$ 000101 $\rightarrow$ 010001, which correspond to $\mathcal{C}_3(4)$, $\mathcal{C}_3(6)$,$\mathcal{C}_3(6)$ and $\mathcal{C}_3(8)$, respectively.

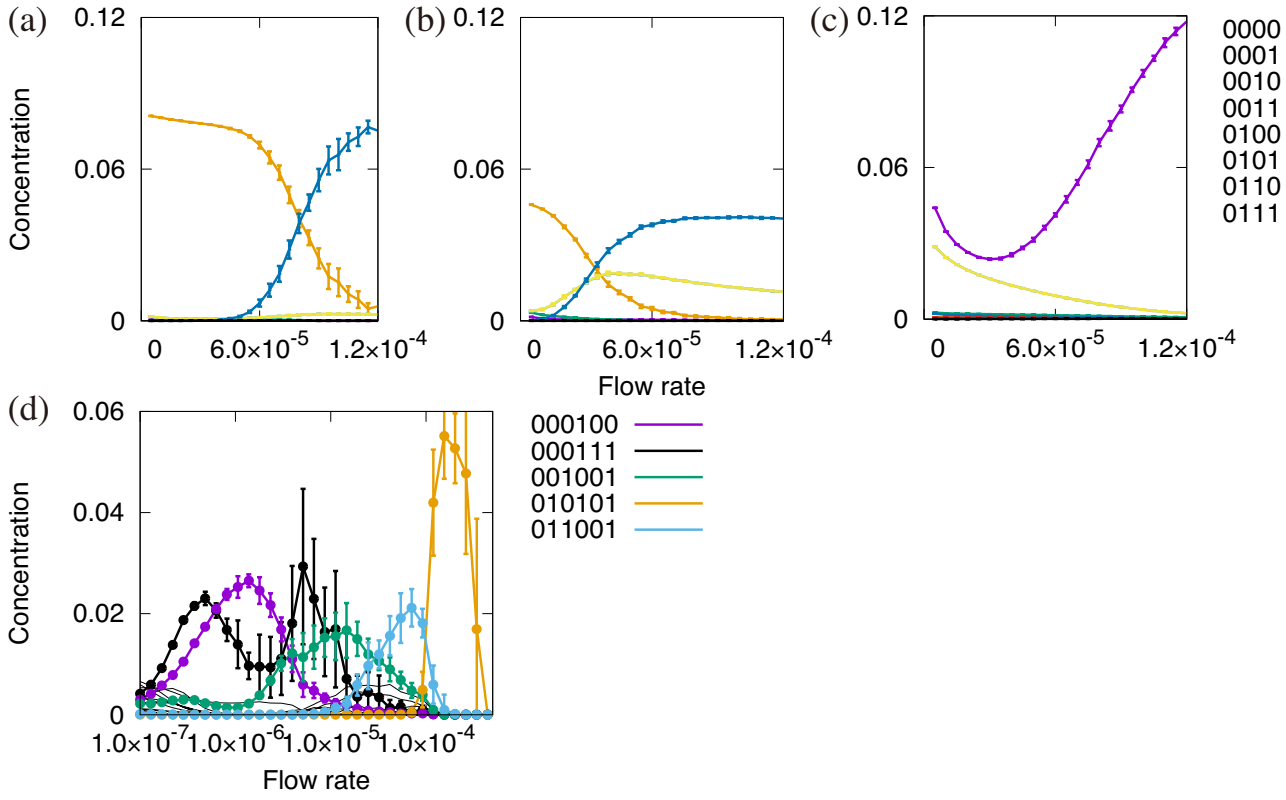

Figure 3.12. Average concentration of each template sequence in the stationary state in the case of bias between monomer species. The horizontal axis is the flow rate in chemostat $f(=d)$. $L = 4, V = 400, f/d = 1.\epsilon = 10^{-4}, f_0 : f_1 = 0.6 : 0.4(a), 0.7 : 0.3(b), 0.9 : 0.1(c)$ and $L = 6, f_0 : f_1 = 0.525 : 0.475$ (d). The most dominant sequence is switched stepwise with the decrease in flow rate as $011001 \to 001001 \to 000111 \to 000100$, which correspond to $\mathcal{C}_3(4)$, $\mathcal{C}_3(6)$,$\mathcal{C}_3(6)$ and $\mathcal{C}_3(8)$, respectively.

between the two flow rates is not too large, qualitatively, the same result is obtained as in the main text Fig. 3.3. This is because in our model, the templates replicate as a complementary pair between 0 and 1, in contrast with the model in ref. [124]. The same number of monomers (0 and 1) is consumed to make each complementary pair of polymers, so that complementary polymers (e.g., 0000 and 1111) take the same concentrations, irrespective, of the monomer concentrations. Thus, the bias in monomer cannot select one of the complementary molecules. As the bias increases, however, the synthesis of templates is slowed down because the minority monomer (i.e. monomer 1) is deficient. Then, the complementary pair hardly exists (i.e. an attractor at which a specific pair of templates exists is unstable) and the catalytic reaction stops. In this case, only the polymer consisting of the majority monomer (i.e. 0000) is synthesized by spontaneous (template independent) ligation reactions.

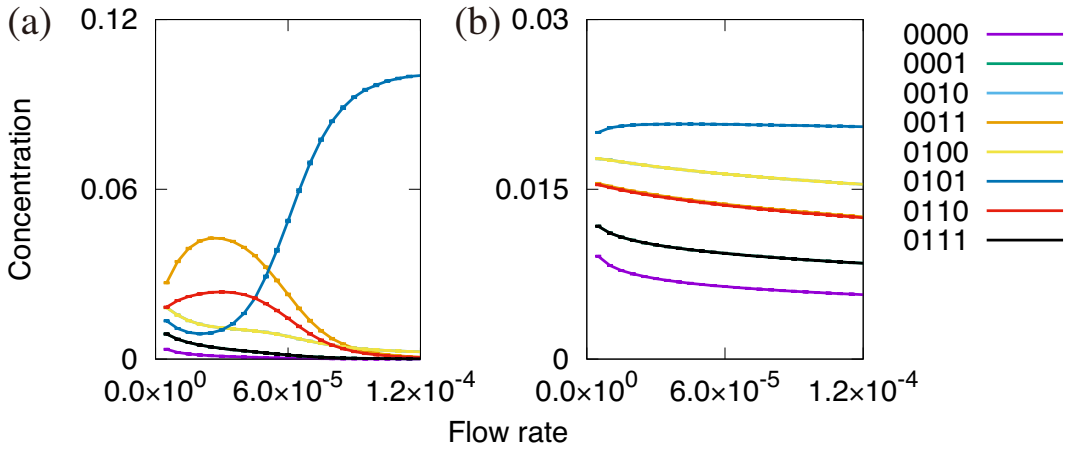

Figure 3.13. Flow rate dependence of average concentrations of templates in the stationary state. The horizontal axis is the flow rate in chemostat $f(=d)$. We set $L = 4, V = 400, \epsilon = 1.0 \times 10^{-3}$ (a) and $\epsilon = 1.0 \times 10^{-2}$ (b).

(v) Change in the spontaneous ligation rate ϵ

We also studied dependence on the spontaneous ligation rate ϵ. The flow rate dependence of the average concentration of template sequences is plotted in Fig. 3.13, as in the main text Fig. 3.3. As shown, the result in Fig. 3.3 (i.e. the dominance of sequence 0101 in the fast flow region, and 0011 in the slow flow region.) is robust against the increase in ϵ up to $\sim 2.0 \times 10^{-3}$. If ϵ is beyond this threshold, the attractor with a dominance of a certain template disappears and the template species just coexist (similar to the error catastrophe [40]), where the flow rate dependence is not observed.

3.6 Summary and Discussion

To sum up, 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) preference of a complex sequence to have a variety of pathways to avoid the deficiency of sub-part shorter polymers.

(I) the finiteness of the system size provides stochasticity (noise) in reaction, which gives rise to switches among attractors with different dominant sequences. Attractors with a higher growth rate have larger stability against noise (see for example Fig. 3.6), 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 (Spontaneous reaction that undergoes without templates plays the 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. see also [64, 125] for attractor selection of a higher cellular growth, and [119, 62, 88] for relevance of noise to chemical evolution).

(II) when the flow rate is limited, the polymerization of 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 [84] 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. (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.) Of course, this is one possible definition, whereas a related definition is adopted to compare the subsequences in the genome in bioinformatics [126, 127, 128, 129, 130, 131]. The relationship between sequence complexity and structural or functional property of polymers has been discussed in several studies [132]. 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 origins 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 origins of life and for the laboratory construction of prebiotic self-replication systems.

3.7 Appendix

3.7.1 Detailed analysis in case of tetramer templates

Derivation of dynamics of two templates

Considering the switch between the dominant sequences of 0010(1101) and 0101(1010), as mentioned in the main text, the rate equation is reduced to the concentrations of these templates and the corresponding trimers, dimers, and monomers.

$$\begin{aligned}
\dot{x}_a &= \frac{1}{2}x_m^2 x_A - \frac{1}{2}x_m x_a x_A - d x_a \\
\dot{x}_b &= \frac{1}{2}x_m^2(2x_A + 2x_B) - \frac{1}{2}x_m x_b(3x_A + 4x_B) - d x_b \\
\dot{x}_{a'} &= \frac{1}{2}x_m(x_a + x_b)x_A - \frac{1}{2}x_m x_{a'} x_A - d x_{a'} \\
\dot{x}_{b'} &= \frac{1}{2}x_m 2x_b(x_A + 2x_B) - \frac{1}{2}x_m(x_A + 2x_B)x_{b'} - d x_{b'}
\end{aligned}$$

$$\dot{x}_A = \tfrac{1}{2}x_m(x_{a'} + x_{b'})x_A \quad -d x_A \tag{3.4}$$

$$\dot{x}_B = \quad x_m x_{b'} x_B \quad -d x_B \tag{3.5}$$

Here, x_m represents the concentration of 0 or 1, x_a is 00 or 11, x_b is 01 or 10, $x_{a'}$ is 001 or 110, $x_{b'}$ is 010 or 101 , x_A is 0010 or 1101, and x_B is 0101 or 1010. Note that the respective complementary pair has the same concentration. The reaction coefficient in the above equations is proportional to the number of corresponding reactions. For example, 00 (a) is produced in a single reaction $0 + 0 + 0010$, and is consumed in the reaction $00 + 1 + 0010$ by the template 0010 (A). On the other hand, 01 (b) is produced in four reactions $0 + 1 + 0010$, $0 + 1 + 1101$, $0 + 1 + 0101$, and $0 + 1 + 1010$ (see main text Fig. 3.5(b)), where B and A are used in two reactions, respectively, and consumed in reactions $0 + 01 + 0010$, $01 + 0 + 0010$, $1 + 01 + 1101$, $01 + 0 + 0101$, $01 + 0 + 1010$, $1 + 01 + 0101$, and $1 + 01 + 1010$, with three reactions using A and four reactions using B.

The concentrations of a, b, a', b' and m are adiabatically eliminated, and two-variable dynamics just for x_A and x_B are obtained as Eq. 3.4 and 3.5 with

$$\begin{aligned}
x_a &= \frac{x_A}{x_m x_A + 2d}x_m^2, \quad x_b = \frac{(2x_A + 2x_B)}{x_m(3x_A + 4x_B) + 2d}x_m^2, \\
x_{a'} &= \frac{x_m x_A(x_a + x_b)}{x_m x_A + 2d}, \quad x_{b'} = \frac{x_m(x_A + 2x_B)(x_b + x_b)}{x_m(x_A + 2x_B) + 2d}.
\end{aligned} \tag{3.6}$$

Using them, the explicit form of r_A and r_B can be written as

$$\begin{aligned}
r_A &= \frac{1}{2}x_m^4\left(\frac{x_A}{x_m x_A + 2d}\left(\frac{x_A}{x_m x_A + 2d} + \frac{(2x_A + 2x_B)}{x_m(3x_A + 4x_B) + 2d}\right)\right. \\
&\quad \left. + 2\frac{(x_A + 2x_B)}{x_m(x_A + 2x_B) + 2d}\frac{(2x_A + 2x_B)}{x_m(3x_A + 4x_B) + 2d}\right), \\
r_B &= 2x_m^4\frac{(x_A + 2x_B)}{x_m(x_A + 2x_B) + 2d}\frac{(2x_A + 2x_B)}{x_m(3x_A + 4x_B) + 2d}.
\end{aligned} \tag{3.7}$$

(Note that x_m is also the function of x_A and x_B but cannot be solved explicitly.)

Derivation of strength of attraction near the fixed point

The Jacobian of the rate equation at a fixed point corresponding to the A-abundant attractor is given by

$$J_A = \begin{pmatrix} \frac{\partial r_A}{\partial x_A}(x_A^*, 0)x_A^* & \frac{\partial r_A}{\partial x_A}(x_A^*, 0)x_A^* \\ 0 & r_B(x_A^*, 0) - d \end{pmatrix}. \tag{3.8}$$

Here, r_A and r_B are the replication rates of A and B, respectively, and $r_A(x_A, x_B) = \frac{1}{2}x_m(x_{a'} + x_{b'})$ and $r_B(x_A, x_B) = x_m x_{b'}$. The eigenvalues of this Jacobian are $\frac{\partial r_A}{\partial x_A}(x_A^*, 0)x_A^*$ and $r_B(x_A^*, 0) - d$, and the maximum eigenvalue of the Jacobian is given by $r_B(x_A^*, 0) - r_A(x_A^*, 0)$, considering that $r_A(x_A*, x_B^*) - d = 0$. Therefore, the replication rates of A and B, i.e., r_A and r_B, determine the maximal eigenvalue of the Jacobian.

Flow rate dependence of concentration of shorter polymers and relation to reaction pathways

For a large $f(= d)$, $x_a = x_A x_m^2/(2d)$ and $x_b = (2x_A + 2x_B)x_m^2/(2d)$, so that $x_a < x_b$. The subsequence b is produced to a greater extent because b is shared by both A and B, while the consumption of b is irrelevant in the fast dilution case (in Eq. 3.6, the dilution term d in the denominator is much larger than the other term). If the supply of dimers is fast, the replication speed of the simple sequence B is higher, as the reactions are concentrated on those using b. Using this, the replication rates r_A and r_B become

$$\begin{aligned} r_A(x_A, x_B) &= \frac{x_m^4}{4d^2}(7x_A^2 + 14x_A x_B + 8x_B^2) \\ r_B(x_A, x_B) &= \frac{x_m^4}{d^2}(2x_A^2 + 6x_A x_B + 4x_B^2). \end{aligned} \tag{3.9}$$

The attraction speed near the fixed point A is given by $r_A(x_A^*, 0) - r_B(x_A^*, 0)(= -\frac{x_m^4}{4d^2}{x_A^*}^2) < 0$ and that near the fixed point B is given by $r_B(0, x_B^*) - r_A(0, x_B^*)(= 2\frac{x_m^4}{d^2}{x_B^*}^2) > 0$.

On the other hand, when $f(= m_0 d)$ is small, $x_a = x_m$ and $x_b = \frac{2x_A+2x_B}{3x_A+4x_B}x_m$, so that $x_a > x_b$. The replication rates are

$$\begin{aligned} r_A(x_A, x_B) &= x_m^2(1 + 3\frac{2x_A + 2x_B}{3x_A + 4x_B}) \\ r_B(x_A, x_B) &= x_m^2(4\frac{2x_A + 2x_B}{3x_A + 4x_B}). \end{aligned} \tag{3.10}$$

When f is small, the attraction speed toward the fixed point A is given by $r_A(x_A^*, 0) - r_B(x_A^*, 0)(= \frac{8}{3}x_m^2) > 0$ and that toward the fixed point B is given by $r_B(0, x_B^*) - r_A(0, x_B^*)(= -\frac{1}{2}x_m^2) < 0$.

In summary, the attraction to the fixed point A is faster (slower) than that to B when f is small (large).

3.7.2 Reduced description of the model

Since the original model requires excessive computational time, we examined the selection in a system size with larger L using the reduced description. Here, the reduction scheme is as follows.

In the reduced model, we consider only the transition between the attractors, in which a certain pair of templates is dominant. During the transition, we assume that the number of templates other than the pair is zero. This is guaranteed when the spontaneous ligation rate, ϵ, is small because a new template will appear only by a spontaneous reaction. For template length L, there are 2^{L-1} attractors. We estimate the transition probabilities, $T_{A\to B}$, between A-abundant and B-abundant attractors by noise.

First, we consider the dynamics of concentration of a pair of templates A and B. By assuming that the dynamics of concentration of the shorter polymer is sufficiently faster than that for the template, the concentrations of polymers shorter than $L-1$ are adiabatically eliminated and determined as functions of the concentration of A and B, i.e., x_A and x_B, so that

$$\begin{aligned}\dot{x}_A =& r_A(x_A, x_B)x_A - dx_A \\ \dot{x}_B =& r_B(x_A, x_B)x_B - dx_B.\end{aligned} \tag{3.11}$$

This equation has two fixed points $x_A = x_A^*, x_B = 0$ and $x_A = 0, x_B = x_B^*$, where $r_A(x_A^*, 0) = d$ and $r_B(0, x_B^*) = d$, corresponding to the A-abundant and B-abundant attractors, respectively.

Then, the corresponding Langevin equations are given by

$$\begin{aligned}\dot{x}_A &= F_A(x_A, x_B) + \sqrt{G_A(x_A, x_B)/V}\eta_A(t) \\ \dot{x}_B &= F_B(x_A, x_B) + \sqrt{G_B(x_A, x_B)/V}\eta_B(t),\end{aligned} \tag{3.12}$$

where $\eta_i(t)$ is the Gaussian white noise with $\langle \eta_i(t)\eta_j(t')\rangle = \delta_{ij}\delta(t-t')$ and $F_A = r_A(x_A, x_B)x_A - dx_A$, $G_A = r_A(x_A, x_B)x_A + dx_A$.

We assume a one-dimensional slow manifold, in which the dynamics connecting the two attractors are restricted on the straight line between the fixed points A and B. We also assume that the replication rate, r_A, and degradation rate, d, are constant along the line, so that $G_A = 2x_A d$ and $G_B = 2x_B d$.

Since we assume that the dynamics are restricted to the one-dimensional line, the probability that the system escapes from attractor A and enters B is calculated by the Kramers formulae, as follows (for example, [54, 55]):

$$T_{A\to B} \propto \frac{1}{G(x)} \exp(2V \int^x \frac{F(x')}{G(x')}dx'), \tag{3.13}$$

where the integral $\int^x$ is carried over from fixed point A to point x along the straight line connecting the fixed points A and B, and x is the value at which the integral takes the maximum value.

Using the approximation $x_A^* \sim x_B^*$, we can now obtain the transition probability explicitly. $T_{A\to B} = \exp(-V\Delta_{A\to B})$, where $\Delta_{A\to B} = \frac{1}{\sqrt{2}d} \max_x \int_0^x (r_A - r_B)\, dx'$. $\Delta_{A\to B}$ is considered to be the highest barrier in the quasi-potential of the dynamics of x_A and x_B along the line $x_A/x_A^* + x_B/x_B^* = 1$. (To calculate r_A and r_B explicitly, we also used $x_A^* \sim f/(2dL)$ by assuming that almost all the monomers (f/d) are included in the templates, and the free monomer concentration $x_m \sim \sqrt{f}$, as derived from Eq. 1 presented in the main text, by assuming that the concentrations of polymers other than the templates are sufficiently low.)

We calculated all the possible transitions between the attractors and constructed the transition matrix T. We assumed that transition occurs only between the pairs of L-mer templates that share the $(L-1)$-mer subsequences. By calculating the eigenvector with the maximum eigenvalue of the transition matrix T, we could obtain the stationary distribution of the model.

We calculated the residence probability of the attractors in the stationary distribution for $L = 4, 8, 12$, as shown in Fig. 3.14, derived from the reduced model.

Although the obtained concentration of the template species differs quantitatively from that obtained by the original chemical reaction model for a small L, the stepwise increase in complexity in the dominant sequence with a decrease in flow rate was reproduced. In other words, at the highest flow rate, the periodic sequence (010101..., $\mathcal{C}_2(2)$) is selected. Here, we also examined the case with a longer L, e.g. $L = 12$. As shown in Fig. 3.14(c), an increase in complexity was seen with a decrease in f.

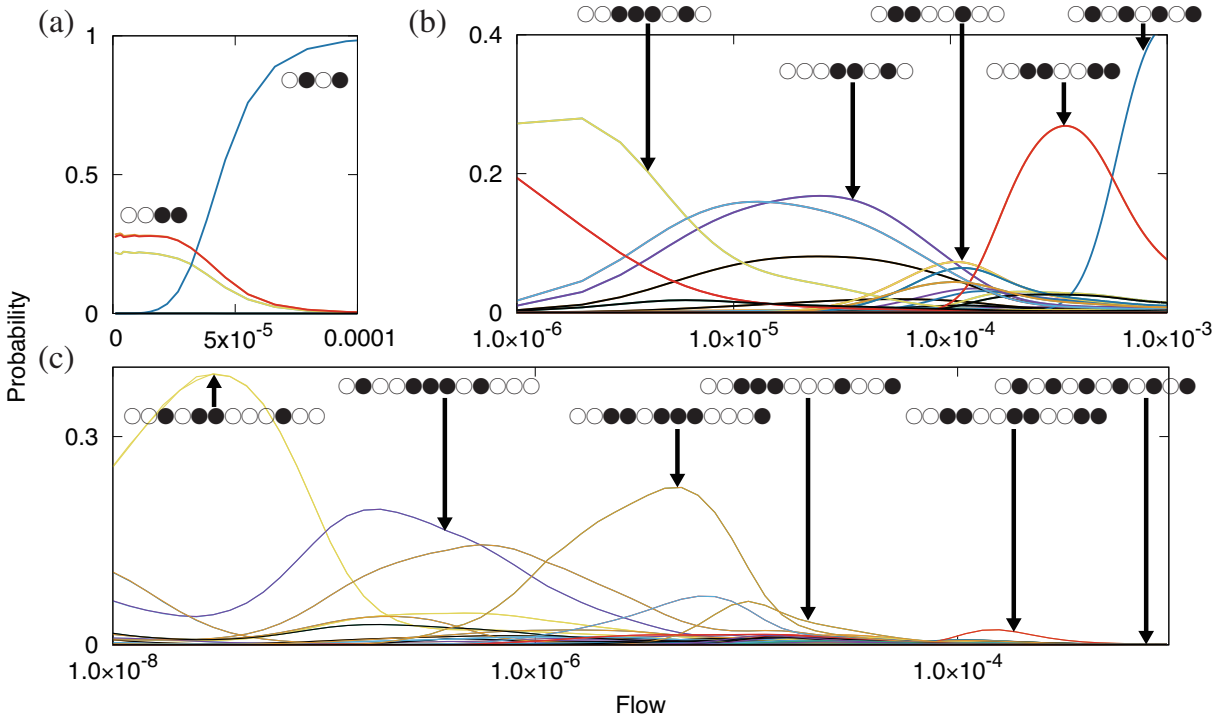

Figure 3.14. Flow rate dependence of stationary probability distribution of attractors with dominance of a pair of templates. Systems with $L = 4$(a), $L = 8$(b), $L = 12$(c) are plotted as a function of flow rate. $V/\sqrt{2} = 50$. Representative sequences are drawn in the figure (black and white circles indicate monomer type 0 and 1, respectively).

Chapter 4

The proto kinetic proofreading in the polymerization process of information polymer

4.1 Introduction

How to prevent the error catastrophe in replication of template polymer is one of the largest problems in the origins of life study. As discussed by Eigen, simple replication system of a polymer with exponential growth shows error catastrophe (see section 1.2.1). The fittest (or functional) sequence cannot keep high population fraction, and is replaced by non fittest ones, due to the accumulation of errors, which is inevitable in thermal polymerization reaction. Here, we study "error catastrophe" problem by using the dynamical model of polymerization. First, we show that the population growth of templates exhibits super- or sub-exponential growth depending on the kinetic parameters. Then, we show that in the super-exponential region, the kinetic error correction mechanism akin to kinetic proofreading works, to prevent the error catastrophe, and the longer template has higher error tolerance. Since this does not require specific design, this may work as an error correction mechanism in the prebiotic world.

4.2 Model

Here we consider the self-replicating template polymer. In the model, we consider the polymerization process and dissociation and association between templates and polymers (Fig.4.1);

$$X_{(l,s)} + X_{(L,t)} \underset{\hat{k}_b(l,s,L,t)}{\overset{k_f}{\rightleftharpoons}} C_{(l,s),(L,t)}, \tag{4-a}$$

$$X_{(1,m)} + C_{(l,s),(L,t)} \underset{\delta r_t}{\overset{r_t \nu(l,s,L,t)}{\rightleftharpoons}} C_{(l+1,sm),(L,t)}, \tag{4-b}$$

where $X_{(l,s)}$ is a polymer with length l and sequence $s \in \{0,1\}^l$, and $C_{(i,s),(L,t)}$ is a complex between a polymer $X_{(l,s)}$ and a template polymer $X_{(L,t)}$. Here for simplicity, only the polymers with length L are assumed to work as templates. The reactions in the first line (4-a) are formation and dissociation of complex between a template $X_{(L,t)}$ and a substrate $X_{(l,s)}$ polymer, and k_f and $\hat{k}_b(i,s,L,t)$ are the rate of these reactions, respectively. The dissociation rate $\hat{k}_b(i,s,L,t)$ depends on the numbers of correct and incorrect base pairs between the sequences of the substrate and the template, which form a complex $C_{(i,s),(L,t)}$. In this model, we assume a monomer in the substrate polymer and one in the template at the same position form a base pair. Further, the pair between different monomer types (i.e., 0 and 1) is assumed to be a correct one, whereas that with the same types (i.e., 0 and 0 or 1 and 1) is incorrect. We assume that energy of the incorrect base pair is higher than the correct one with the difference ϵ. Therefore the dissociation rate $\hat{k}_b(i,s,L,t)$ is given by

$$\hat{k}_b(i,s,L,t) = k_b \exp(n\epsilon),$$

where $n(\leq l)$ is the number of incorrect pairs between a template $X_{(L,t)}$ and a substrate $X_{(l,s)}$. Thus, the more mismatch pair in a complex exists between a template and a substrate, the dissociation of them occurs faster. For example, since the complex of 0100 and 101 has no mismatch base pair, the dissociation rate of them is $\hat{k}_b(3,101,4,0100) = k_b$. While, $\hat{k}_b(3,111,4,0100) = k_b e^{\epsilon}$, since the complex of 0100 and 111 has one mismatch base pair.

The reaction in the second line (4-b) is the ligation between a monomer and a polymer which forms a complex with a template. In this model, a polymer which does not form a complex with a template cannot elongate. Since there are two types of monomers, there are two types of ligation reaction; with a correct base pair and an incorrect one. If a base pair of the monomer to be ligated and the one at the corresponding position in the template $X_{(l,s)}$ are correct ones, the ligation rate is r_t, and otherwise, the rate is $r_t e^{-\epsilon}$. We also assume that the ligation rate r_t depends on the sequence of the template $X_{(L,t)}$, and the rate is the same as its complemental sequence. For example, in the case with a substrate 101 which forms a complex with a template 0100, a monomer 0 ligates at the rate r_{0100} and a monomer 1 at the rate $r_{0100}e^{-\epsilon}$. Note that $\mu = 1/(1+\exp(\epsilon))$ is the error rate at each monomer addition in the template replication.[1]

In the model, we also assume the constant supply of free monomers 0, 1 at rate $f/2$, respectively and the constant dilution of polymers at rate ϕ (chemostat condition);

$$\emptyset \xrightarrow{f/2} X_{(1,m)}, \qquad X_{(l,s)} \xrightarrow{\phi} \emptyset, \qquad C_{(l,s),(L,t)} \xrightarrow{\phi} \emptyset,$$

where m represent monomer types 0 or 1. Thus total monomer concentration is given by $x^t_m = \sum_l \sum_{s\in l} \left[l x_{(l,s)} + (L+l) \sum_{t\in L} c_{(l,s),(L,t)} \right]$, where $\sum_{s\in l}$ is the summation for all of the sequences of length l, which is kept as constant value f/ϕ.

[1]Here, we consider only the ligation of a monomer to a right side of a polymer. Therefore, unlike the model in the previous chapter, there is no difference in the diversity of reaction pathways in synthesizing templates with different sequences.

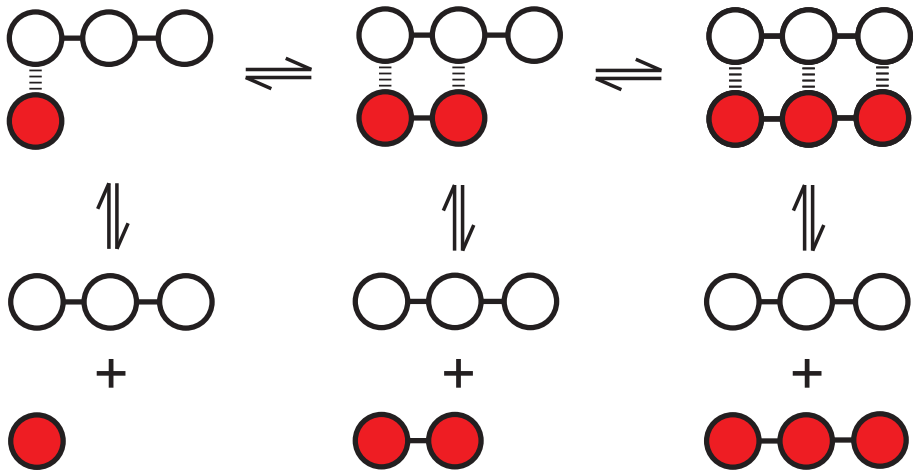

Figure 4.1. Schematic diagram of the polymerization model. In the upper row, a complex between a trimer template and a monomer, a dimer and a trimer with complemental monomer sequence are drawn. Right arrows between them represent ligation reactions between a substrate polymer and a monomer and left arrows reverse of these reactions. There are two types of monomers (red and white circle), and the dashed line represents a base pair between them. Up and down arrows represent complex formation with a trimer template and a substrate.

The rate equation of the concentration of polymers and complexes in this chemical reaction process (assume infinite system size) is,

$$\begin{aligned}
\dot{x}_{1,m} =& f/2 - k_f \sum_{t\in L} x_{1,m}x_{L,t} + \sum_{t\in L} \hat{k}_b(1,m,L,t)c_{1,m,L,t} \\
&- \sum_{l,s\in l,L,t\in L} r_t \nu(l+1,sm,L,t)c_{l,s,L,t}x_{1,m} - \phi x_{1,m}, \\
\dot{c}_{1,m,L,t} =& \quad k_f x_{1,m}x_{L,t} - \hat{k}_b(1,m,L,t)c_{1,m,L,t} - r_t c_{1,m,L,t}x_{1,m} - \phi c_{1,m,L,t}, \\
\dot{x}_{l,s} =& \quad - k_f \sum_{t\in L} x_{l,s}x_{L,t} + \sum_{t\in L} \hat{k}_b(l,s,L,t)c_{l,s,L,t} - \phi x_{l,s}, \\
\dot{c}_{l,s,L,t} =& \quad k_f x_{l,s}x_{L,t} - \hat{k}_b(l,s,L,t)c_{l,s,L,t} + r_t \nu(l-1,s',L,t)c_{l-1,s',L,t}x_{1,m} \\
&- r_t(1+e^{-\epsilon})c_{l,s,L,t}x_1 - \phi c_{l,s,L,t}, \\
\dot{x}_{L,t} =& \quad - k_f x_{L,t}x_{L,t} + \hat{k}_b(L,t,L,t)c_{L,t,L,t} + \sum_{l,s\in l}(-k_f x_{l,s}x_{L,t} + \hat{k}_b(l,s,L,t)c_{l,s,L,t}) - \phi x_{L,t}, \\
\dot{c}_{L,s,L,t} =& \quad k_f x_{L,s}x_{L,t} - \hat{k}_b(L,s,L,t)c_{L,s,L,t} + r_t \nu(L-1,s',L,t)c_{L-1,s',L,t}x_{1,m} - \phi c_{L,s,L,t},
\end{aligned} \tag{4.1}$$

where s' represents the sequence in which the right end monomer of the sequence s is deleted. Note that the sequence and its complementary sequence have the same concentration at any time. Also note that these reactions satisfy the detailed balance condition, thus in the equilibrium state (i.e. the stationary state when there are no nonequilibrium flow $f = \phi = 0$), the total concentration of a polymer $X_{(l,s)}$ i.e. $x^t_{l,s} = x_{l,s} + \sum_{t\in L} c_{l,s,L,t}$ is the same for all of the sequences with the identical length.

4.3 Result

4.3.1 Competition between two types of templates under no mutation

We first test the competition between two complementary pairs of templates in the condition in which there is no mismatch ligation of a monomer; i.e., the energy of the incorrect base pair ϵ is infinitely large. We assume that these pairs of templates share only monomer species as a substrate. Below, we study the dynamics of the concentration of only one of the complemental template pair, each complemental template has the same concentration.

First, we consider the case in which their elongation speed constants r are same. We also set the length of templates are three. We calculated the dynamics of the concentrations of two templates in Fig. 4.2. As in the figure, there are two types of fixed points: in one fixed point two template pairs coexist, and in the other, only one template pair dominates. The stability of the fixed point depends on the dissociation rate k_b as shown in Fig. 4.3(a). When k_b is low, only the fixed point of co-existence is stable. On the other hand, when k_b is high, fixed point with the

dominance of one of them is stable. In the middle range of k_b, both of them are stable, and the initial condition of concentrations determines the final state. The same phase diagrams are drawn for the case in which the template polymer is longer. As the figure in Fig. 4.3(b), The phase diagram in these cases is qualitatively the same as in the case of $L = 3$, although the parameter region in which the selection of two types of templates shows frequency dependency is extended to the region with lower k_b.

Next, we set the ligation speed constants r to be different between two pairs of templates. In proportion to the speed, the templates are replicated. Hence the speed works as the fitness. Selection of a template sequence is made accordingly. As shown in Fig. 4.4, when k_b is high, both of the fixed points with the dominance of the templates and that with the coexistence are stable. Here, the former has higher fitness than the other. As k_b decreases, the fixed point with the dominance of lower fitness template become unstable first before one with higher fitness. Thus there is a regime where only the fixed point with the dominance of the template with the highest fitness is stable.

To understand these results, here, we reduce the rate equation of the model chemical reaction system and derive the growth rate of the templates concentration. When the ligation rate is much faster than the complex formation and deformation, $k_b \ll k_f \ll r$, the model chemical reaction system can be reduced to,

$$X_1 + X_L \overset{k_f}{\to} C_L$$
$$2X_L \underset{k_b}{\overset{k_f}{\rightleftharpoons}} C_L,$$

where X_1 is a monomer, X_L is a template species and C_L is complex of two of templates. In this case the rate equation of concentration of X_L is approximated as

$$\dot{x}_L^t = x_1\sqrt{2k_f k_b}\sqrt{x_L^t} - \phi x_L^t. \tag{4.2}$$

(detailed calculation is in the last of this chapter) In this rate equation, the growth of template is not proportional to its concentration x_L^t but with its square-root. This is the reason why the templates coexist in the low k_b regime, as discussed in section 1.2.1.

On the other hand, when the association and dissociation rates are sufficiently faster than the ligation rate, $k_b \gg k_f \gg r$, the model can be reduced to,

$$X_1 + X_L \overset{rK}{\to} X_2 + X_L \overset{rK}{\to} \ldots \overset{rK}{\to} 2X_L,$$

where $K = k_f/k_b$. If we also assume the dilution rate is sufficiently faster than the ligation reaction i.e., $\phi \gg rKx_1$, we obtain

$$\dot{x}_L^t = (r^{L-1}K^{L-1}{x_1^t}^L/\phi^{L-2}){x_L^t}^{L-1} - \phi x_L^t, \tag{4.3}$$

The exponent of this growth rate equation is proportional to the number of steps in the polymerization of templates, i.e. the length of the template L. Intuitively,

this is because the template is required at multiple steps of ligation reactions as a catalyst. Since the exponent of the growth rate of templates is larger than 1 as discussed in section 1.2.1. This is the reason why the sequence that is dominant initially is selected for larger k_b.

To estimate the parameter k_b that separates the regions of sub- and super-exponential growth, we consider the reduced form of the chemical reaction, with the case $L = 3$

$$X_1 + X_3 \overset{rK}{\to} X_2 + X_3 \overset{rK}{\to} C_3,$$
$$2X_3 \underset{k_b}{\overset{k_f}{\rightleftharpoons}} C_3.$$

Here, the formation of the complex between a template and substrate polymers shorter than the template is assumed to be sufficiently faster than the other reactions.

In this case, the growth rate of the template is given by

$$\dot{x}_3^t = \alpha x_1 \phi \frac{(\sqrt{1+\beta x_3^t} - 1)^2}{(\sqrt{1+\beta x_3^t} - 1) + \alpha}, \tag{4.4}$$

where $\alpha = \frac{4k_f\phi}{rKx_1(\phi+k_b)}$ and $\beta = \frac{8k_fk_b}{(k_b+\phi)^2}$. From this description, the condition that the template grows super-exponentially is estimated as

$$1 \gg \beta x_3^t. \tag{4.5}$$

This condition is satisfied when k_b is large, and k_f is small. According to Eq. 4.5, the critical value for the super-exponential growth depends on the total concentration of the template. This explains why there exist two attractors one of which shows coexistence of templates and the other represents the survival of only one template, in the middle region of k_b in Fig.4.3. This formula can also be applied for longer templates. If the total concentration of the monomers (i.e., the supply rate of a monomer f and the diffusion rate of ϕ) is fixed, the coexistence of longer templates tends to be harder because they have low concentrations.

4.3.2 Replication of templates with the error

Next, we analyze the replication of template polymer under error and explain the error tolerance of their sequence information (i.e., the energy difference between the correct and incorrect base pairs ϵ is finite). Here we assume the pair of template sequences with only one type of monomers (i.e., 000... and 111...) as the master sequences and set the replication rate r_0 of the master sequence as $r_0 = 1.0$, and these of all the other r_1 as $r_1 = 0.1$. For this parameters, there is only the fixed point of the dominance of the master sequences. (i.e., the survival of the fittest).

We calculated the dependency of the fraction of the master sequence for all of the template sequences with the error rate μ. The fraction represents the precision

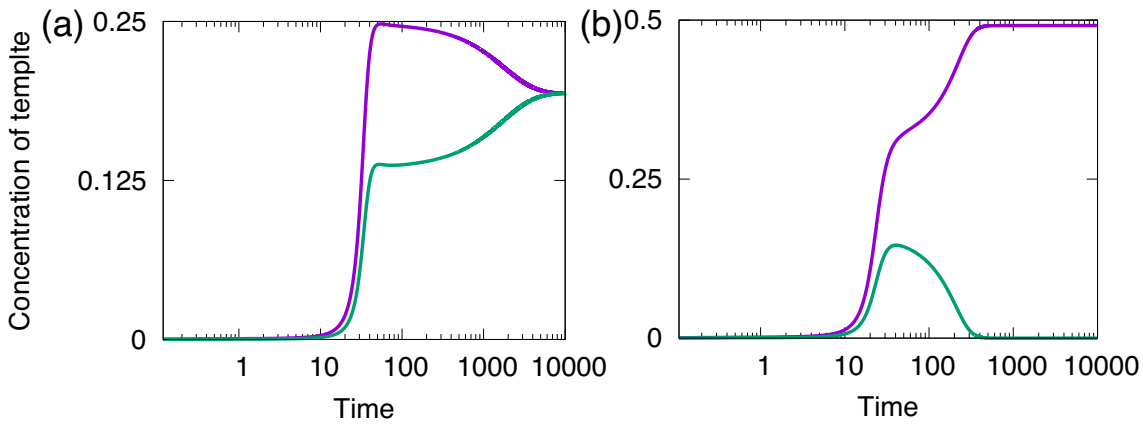

Figure 4.2. Time series of the total concentrations of two template species A (green), B(blue), x_A^{tot}, x_B^{tot}. In initial condition the concentration of template A is slightly larger than B. We set $\phi = 0.1, f = 1.0$, $k_b = 1$ (a) and $k_b = 3$ (b). The horizontal axis is time in log scale. For low k_b (left), the two templates eventually converge to same concentration at the stationary state, and for high k_b the template A dominates the system.

of information replicated. Naturally, the fraction decreases monotonically with the increase of μ, as in Eigen's quasispecies model. As in Fig. 4.5(a), with the increase in the dissociation rate k_b, the decline of the fraction with the increase of μ is moderate than in the quasispecies model with the same length of templates and same fitness distribution. This is consistent with the discussion in the previous section that the growth of the templates is super-exponential for high k_b. Thus the initially abundant master sequences are better preserved and more dominated.

Next, we plotted the same figure with various length L of the template. (Fig. 4.5(b)) Surprisingly, the decrease of the fraction of the master sequence with the increase of error μ is slower with the increase of L. This result is in a sharp contrast with that in the quasispecies model, in which the decline of the fraction is sharper with the increase of L, as studied on error catastrophe (section 1.2.1). This is because the exponent of the growth rate of templates is larger as their length is longer, and the master sequence is more dominantly selected.

There is a tradeoff relationship between the accuracy and the speed of replication of the templates. In Fig. 4.6, how the fraction of the master sequence among all of the template sequences and the total concentration of the master sequence depends on the dissociation rate k_b is plotted. The fraction increases with the increase of k_b, whereas the yield decreases in a high k_b region. Although the increase of the fraction is monotonous for all regions of k_b, there is the optimal value for k_b which gives the maximum of the yield.

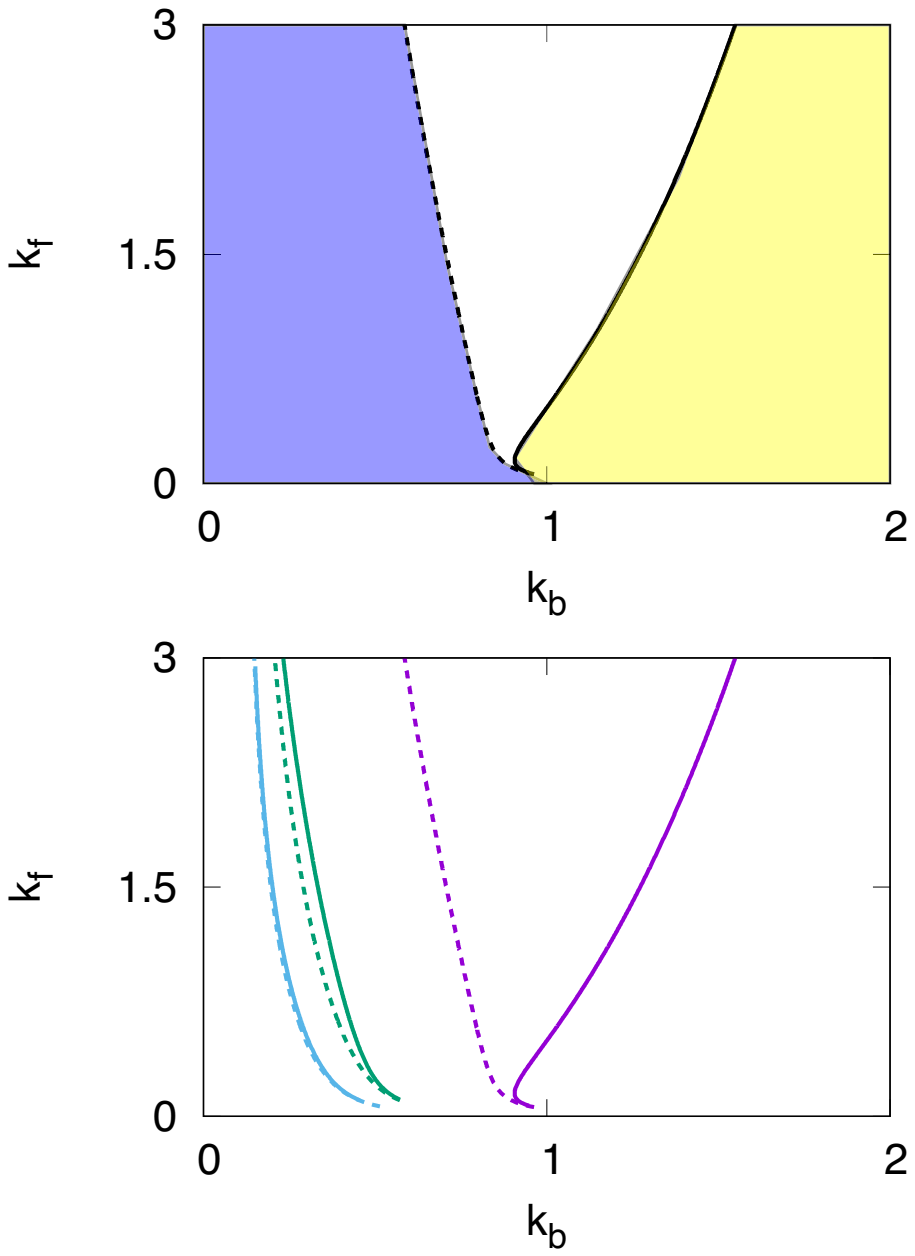

Figure 4.3. (a)Phase diagram of stability of the fixed points. Horizontal axis is dissociation rate k_b, and the vertical is association rate k_f. (I)Only the fixed point in which two templates coexist is stable (blue area). (II)Only the fixed point in which one template dominates is stable (yellow area). (III)Both types of fixed points are stable (white area). (b)The same diagram for $L = 3$ (blue), $L = 4$ (green) and $L = 5$ (light blue). At the left side of the solid line the fixed point with the coexistence of the two templates is stable, whereas the other side it is unstable. At the right side of the dashed line, the fixed point with the survival of the only one is stable. Thus, there are three phases.

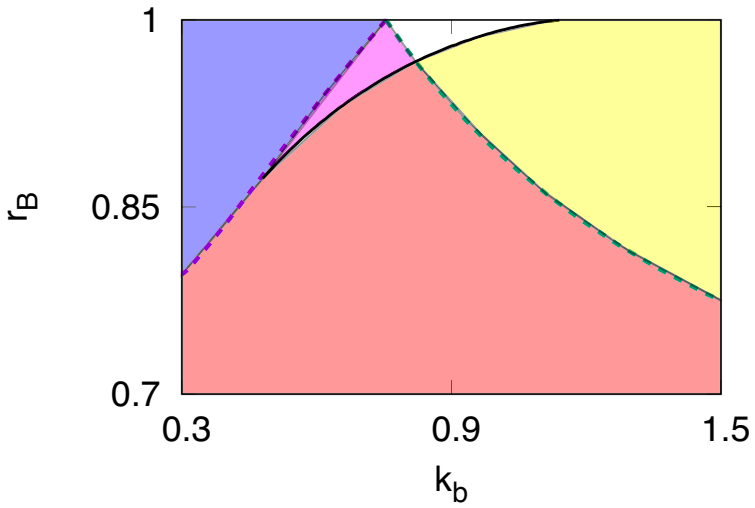

Figure 4.4. Phase diagram of stability of the fixed points for two templates with different fitness. The horizontal axis is k_b and the vertical axis is the fitness of the second template pair B, r_B. We set $f = 1, \phi = 0.1, L = 3$ and the fitness of the template pair A as $r_A = 1$. There are three types of attractors; (i) in which template A and B coexist, (ii) template A dominates, and (iii) template B dominates. Colors of each area represent the stability of these attractors; only attractor (i) is stable (blue), (ii) and (iii) are stable (yellow), (i) and (ii) are stable (purple), and all of the attractors are stable (white). Solid line represents the boundary of stability for the fixed point with coexistence of two templates. Blue and orange dashed line represent the boundary of stability for the fixed point with dominance of template A and one with B respectively.

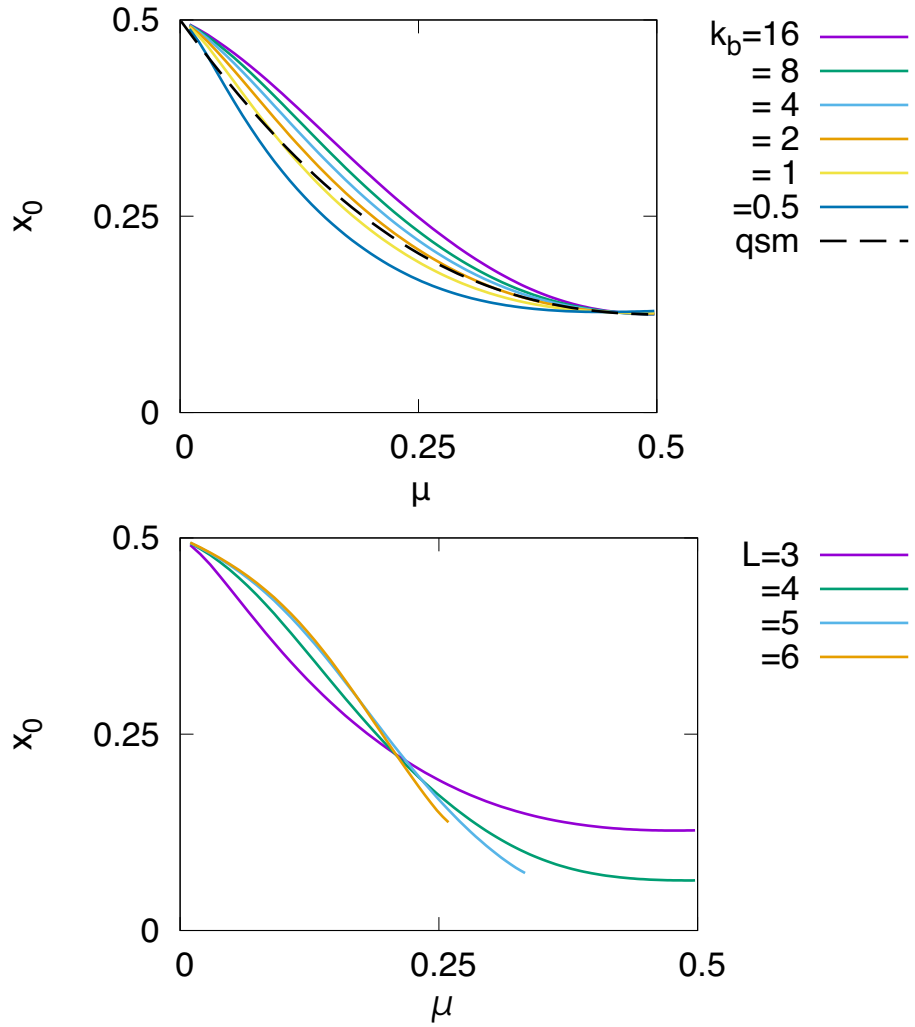

Figure 4.5. Dependence of the fraction of one of the master sequences among all of the template sequences $x_0 = x_{L,0}/\sum_{s\in L} x_{L,s}$ on the error rate μ. We set $f = 1, \phi = 0.1, k_f = 1, L = 3$ and $k_b = 0.5, 1, 2, 4, 8, 16$. Dashed line (qsm) represents the dependence of x_0 on μ in the model in which templates grows exponentially corresponding to the quasispecies model with $L = 3$ (see detail in the last of this chapter.). The same plot for different set up for the length of templates L. ($f = 1, \phi = 0.1, k_f = k_b = 1.$)

4.3.3 Similarity with kinetic proofreading and trade-off between replication speed and accuracy

The polymerization process can be regarded as multi-step error correction of each monomer in the template sequence. After the ligation of i-th monomer to substrate polymer, there are $L-i+1$ step of ligation reactions before completion of template synthesis. Then the relative speeds of each ligation between the templates in which the i-th monomer is 0 and 1, $F_0^{(i)}$ and $F_1^{(i)}$ respectively, are

$$\begin{aligned} F_0^{(i)} &\propto e^{\epsilon}(x_0 r_0 + (x_0^{(i)} - x_0)r_1) + x_1^{(i)} r_1 \\ F_1^{(i)} &\propto x_0 r_0 + (x_0^{(i)} - x_0)r_1 + e^{\epsilon} x_1^{(i)} r_1, \end{aligned} \tag{4.6}$$

where $x_0^{(i)}$ and $x_1^{(i)}$ are the fractions of template sequence such that the i-th monomer is 0 or 1, so that $x_0^{(i)} + x_1^{(i)} = 1$. Then the fraction of one of the master sequence with all-0 is $x_0 = \prod_i x_0^{(i)}$. Here, we ignore the other complemental master sequence and sequences similar to it, since they hardly form a complex with sequences considered here for large ϵ.

Thus, relative production rate of templates that i-th monomer is 0 or 1 are derived as

$$\begin{aligned} \dot{x}_0^{(i)} &\propto (F_0^{(i)})^{L+1-i}, \\ \dot{x}_1^{(i)} &\propto (F_1^{(i)})^{L+1-i}. \end{aligned} \tag{4.7}$$

If we assume the dominance of the master sequence ($x_0^{(i)} \sim x_0$, $x_1^{(i)} \sim 0$), the effective error rate at the i-th monomer in replication, $\mu^{(i)}$, is approximated as

$$\frac{(F_1^{(i)})^{L+1-i}}{(F_0^{(i)})^{L+1-i} + (F_1^{(i)})^{L+1-i}} \sim e^{-(L+1-i)\epsilon}, \tag{4.8}$$

since $F_1^{(i)}/F_0^{(i)}$ is approximated as $e^{-\epsilon}$ This is exactly same as minimum error rate that can be achieved in the kinetic proofreading model with $L+1-i$ steps (see section 1.2.4). In this way, in our model, if the error rate at the i-th monomer of the template replication is μ^{L+1-i}, the error threshold for information persistence takes a finite value even for infinitely long templates, in contrast with the quasispecies model.

Also note that (relative) growth rates of $x_0^{(i)}$ and $x_1^{(i)}$ are proportional to ${x_0^{(i)}}^{L+1-i}$, ${x_1^{(i)}}^{L+1-i}$, respectively, for large ϵ. Thus their growth is super-exponential. In this error correcting scheme, the super-exponential growth of a template is inevitable, because to energetically discriminate correct and incorrect base-pairs fast association and dissociation between a substrate and a template is needed.

4.3.4 Optimality for evolution speed

Finally, we study the speed of the evolution in the template system. We calculated the time series from the state in which the dominance of lower fitness template

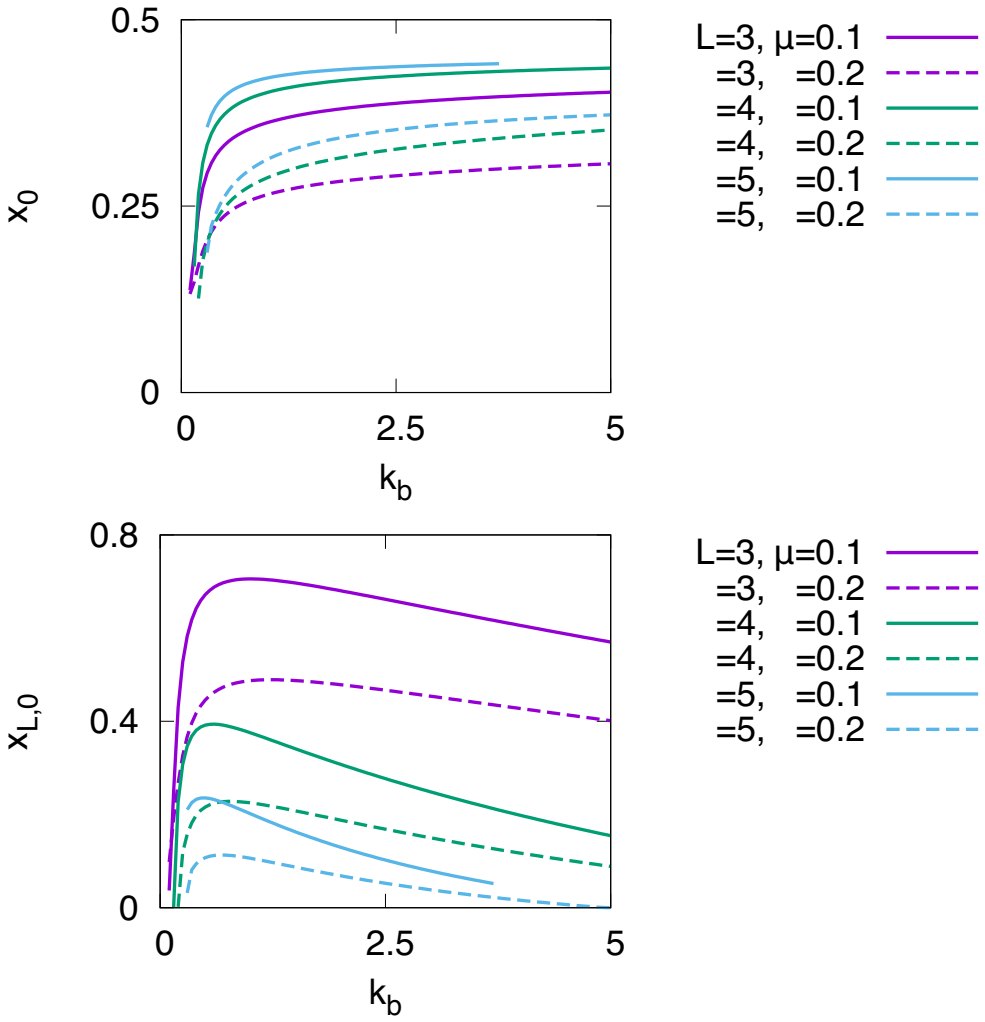

Figure 4.6. (a) Dependance of the fraction of one of the master sequence among all of the templates x_0 on the dissociation rate k_b. (b) Dependence of the total concentration of one of the master sequences $x_{L,0}$ on k_b.

to the state of the dominance of the fittest template. We set the fitness of all 0 and all 1 sequence template (master sequences) as $r_0 = 1.0$, whereas that of the periodic sequence 0101... and 1010... (which are the most far from both of the master sequence in the sequence space.) is set at $r_1 = 0.5$, and the others are set at $r_2 = 0.1$. In this setting considered here, only the fixed point with the dominance of the master sequences is stable. Thus, no matter what the initial condition of the fraction of templates is, the master sequence eventually dominate the system.

We first start from the dominance of the (complemental pair of) periodic sequence and calculate the dynamics of the templates concentrations to converge to the fixed point. (Fig. 4.7(a)) Here we define the speed of the evolution of templates by the transient time until the master sequence becomes dominant. As in Fig. 4.7(b), there exists an optimal k_b value that minimizes the evolution time.

For small k_b, the speed of the template evolution is low, since the selection pressure of the master sequence is weak because the templates grow sub-exponentially. On the other hand, for large k_b the evolution is also slow. This is because the template abundance grows super-exponentially so that initially dominant sequence that is not fittest keeps its dominance over other sequences (see section 1.2.1).

4.4 Summary and Discussion

To sum up, we found in the template polymerization model there are phases of sub- and super-exponential growth corresponding to the growth form discussed in section 1.2.1. In the super-exponential growth phase, the longer template has higher degrees of power in the growth rate. Therefore, in that case, the longer template has higher error tolerance in contrast to Eigen's original model. Here, however, there is a trade-off between the accuracy of information in template sequence and the production speed of the template. Also, there is an optimal point to maximize the evolution speed of templates, in which their growth is exponential.

In this chapter, we reconsider Eigen's theory for the selection of replicators and the error catastrophe in the quasispecies model, from the microscopic elementally polymerization process. In our model, the growth of templates changes from sub to super exponential depending on the kinetic parameter of polymerization as in Eigen's model. On the other hand, unlike Eigen's model of the replicator, the exponent of the growth rate of templates also depends on their concentration at the moment. Further, since long templates require many steps for their synthesis, they have a higher exponent of the growth rate. Accordingly, they are more robust against errors in the information replication compared with the simple exponential growth.

In this chapter, we assume only the long polymer can be a template. Thus our model cannot solve the problem of Spiegelman's monster discussed in section 1 [30, 31, 32]; Since shorter templates have faster replication rates, longer templates are not selected. Instead, we focused on the selection of the sequence with the correct information among a vast range of the sequence space. For example, the condensation of longer polymers by thermal convection in rock pore is proposed

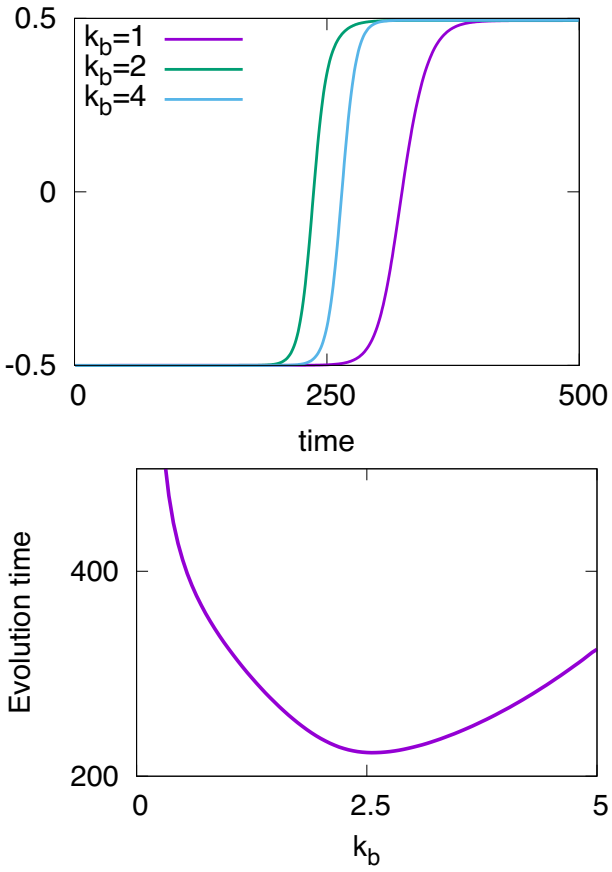

Figure 4.7. (a) Time series in the difference of the fractions between the master sequences (00000) or (11111), x_0 and the sequences with the second highest fitness (01010) or (10101), x_1 for $k_b = 1, 2$, and 4. We set the initial condition in which the fraction of the second fittest is 1. $f = 1, \phi = 0.1, L = 5$. (b) The evolution time from the initial sequences to the master sequence with varying k_b value. The horizontal axis is k_b and the vertical axis is the time until the fraction of the master sequence exceeds the second fittest.

as a solution to the Spiegelman's monster problem [133, 134]. Our model suggests that if there is such a mechanism for selecting longer polymers, the problem in the information inheritance is also settled.

This mechanism to increase the replication accuracy is similar to kinetic proofreading, which is an error correction mechanism proposed by Hopfield [92, 93]. Kinetic proofreading is used in the present cell to reduce the error in the replication by doubling the energy difference between correct and incorrect base-pairs. On the other hand, the proofreading requires more time for synthesis of the polymer. In fact, our model exhibits also the tradeoff between the speed and the accuracy of the production [94, 95, 96, 97]. In order for kinetic proofreading to work, the specific design to couple the synthesis of a molecule with the conversion of the cofactor such as ATP is needed. On the other hand, in the scheme in this chapter, only the requirement is the supply of energy as a monomer and dilution of polymers. Therefore, it can provide a plausible mechanism of error correction in the prebiotic world.

A super-exponentially growing template is advantageous to maintain accurate replication under a high error rate. On the other hand, with exponential growth, both of the growth rate and the evolution speed are moderately fast. Thus, life systems in the early stages would have used error correction mechanism proposed in our model, and then gradually evolved to the regime of exponential growth to satisfy Darwinian evolution.

4.5 Appendix

4.5.1 Derivation of the growth rate of templates

Sub-exponential growth for $k_f \gg k_b$

When association rate between polymers are faster than dissociation,$k_f \gg k_b$, the model chemical reaction system can be reduced to as followings,

$$X_1 + X_L \to C_L \leftrightarrow 2X_L$$

where X_1 is a monomer, X_L is a template species and C_L is double-helix of two of templates.

The rate equation is

$$\begin{aligned} \dot{x_L} &= -2k_f x_L^2 + 2k_b c_L - k_f x_L x_1 + k_b c_1 - \phi x_L, \\ \dot{c_L} &= k_f x_L^2 - k_b c_L + k_f x_L x_1 - \phi c_L. \end{aligned} \tag{4.9}$$

The rate equation for the template polymer $x_L^t = 2c_L + x_L$ is

$$\dot{x_L}^t = \frac{d}{dt}(2c_L + x_L) = k_f x_L x_1 - \phi x_L^{tot}. \tag{4.10}$$

Then if we assume $k_f \ll k_b$, $x_L^t \sim 2c_L$ and $k_b c_L \sim k_f x_L^2$. It can be concluded that

$$\dot{x}_L^t = x_1\sqrt{2k_f k_b}\sqrt{x_L^t} - \phi x_L^t. \tag{4.11}$$

Super-exponential growth for $k_f \ll k_b$

On the other hand, when the dissociation rate is faster than association rate, $k_f \ll k_b$, the model where the template is trimer can be reduced to,

$$X_1 + X_L \to X_2 + X_L \to \cdots \to 2X_L,$$

where $K = \frac{k_f}{k_b}$.

If the complex formation and deformation reactions are much faster than the ligation reaction,$k_b \gg 1$ and $k_f \gg 1$, $c_3 = \frac{k_f}{k_b}x_3x_3, c_2 = \frac{k_f}{k_b}x_3x_2, c_1 = \frac{k_f}{k_b}x_3x_1$, Total tetramer, dimer and monomer concentrations are, $x_3^t = x_3+2c_3+c_2+c_1$, $x_2^t = x_2+c_2$, $x_1^t = x_1 + c_1$. For $k_b \gg k_f$, Then the rate equation is reduced to,

$$\begin{aligned} \dot{x}_1^t &= f - 2rKx_3^t(x_1^t)^2 - rKx_3^t x_2^t x_1^t - \phi x_1^t, \\ \dot{x}_{l+1}^t &= -rKx_1x_L^t x_{l+1}^t + rKx_L^t x_l^t x_1^t - \phi x_l^t, \\ \dot{x}_L^t &= rKx_1^t x_L^t x_{L-1}^t - \phi x_L^t. \end{aligned} \tag{4.12}$$

If we assume the stationary state (or change of dimer's total concentration is much faster than tetramer's),

$$x_{l+1}^* = \frac{rKx_1x_ix_L}{rKx_1x_L + \phi} \sim Kx_1x_ix_L/\phi. \tag{4.13}$$

The approximation can be applied when the speed of dilution is much faster than the speed of ligation reaction, $\phi \ll Kx_1x_L$. As a result, the rate equation for template concentration is

$$\dot{x}_L^t = (r^{L-1}K^{L-1}{x_1^t}^L/\phi^{L-2}){x_L^t}^{L-1} - \phi x_L^t. \tag{4.14}$$

Estimation of the parameter k_b that separates the regions of sub- and super- exponential growth

$$X_1 + X_3 \overset{rK}{\to} X_2 + X_3 \overset{rK}{\to} C_3,$$
$$2X_3 \underset{k_b}{\overset{k_f}{\rightleftharpoons}} C_3.$$

The rate equation is

$$\begin{aligned}
\dot{x}_2^t &= rKx_1^2x_3 - rKx_1x_2x_3 - \phi x_2^t, \\
\dot{x}_3 &= -2k_f x_3^2 + 2k_b c_3 - \phi x_3, \\
\dot{c}_3 &= k_f x_3^2 - k_b c_3 + rKx_1x_2x_3 - \phi c_3.
\end{aligned} \tag{4.15}$$

If we assume $\dot{x}_2^t = 0$,

$$x_2^t = \frac{rKx_1^2x_3}{rKx_1x_3 + \phi}. \tag{4.16}$$

From constraint for conservation of total template concentration, $x_3^t = x_3 + 2c_3$, and $\dot{x}_3 = 0$,

$$x_3 = \frac{k_b + \phi}{4k_f}\left(\sqrt{1 + \beta x_3^t} - 1\right), \tag{4.17}$$

where $\beta = \frac{8k_fk_b}{(k_b+\phi)^2}$.

Therefore, the growth rate of the template is

$$\dot{x}_3^t = rKx_1x_2^tx_3 = \frac{rKx_1^2(k_b+\phi)}{4k_f}\frac{(\sqrt{1+\beta x_3^t}-1)^2}{(\sqrt{1+\beta x_3^t}-1)+\alpha}, \tag{4.18}$$

where $\alpha = \frac{4k_f\phi}{rKx_1(\phi+k_b)}$.

When $1 \ll \beta x_3^t$, from the approximation $\sqrt{1+\beta x_3^t} \sim 1 + \frac{1}{2}\beta x_3^t$, the growth of the template is faster than exponential,

$$\dot{x}_3^t \propto \frac{(\frac{1}{2}\beta x_3^t)^2}{\beta x_3^t + \alpha}. \tag{4.19}$$

When $1 \ll \beta x_3^t$, on the other hand, from the approximation $\sqrt{1+\beta x_3^t} \sim \sqrt{\beta x_3^t}$, the growth of the template is slower than exponential,

$$\dot{x}_3^t \propto \frac{\beta x_3^t}{\sqrt{\beta x_3^t} + \alpha}. \tag{4.20}$$

Therefore, the critical k_b that separates the regions of sub- and super- exponential growth is

$$1 \sim \beta x_3^t \tag{4.21}$$

The polymerization process which can be reduced to the quasispecies model

Here we change the dissociation rate $\hat{k}_b(l, s, L, t)$ as followings:

$$\begin{aligned}\hat{k}_b(l, s, L, t) &= \exp(n\epsilon) \quad (l = 1),\\ &= 0 \quad (l = 2 \dots L - 1),\\ &= \infty \quad (l = L).\end{aligned}$$

The dissociation of a complex between a polymer longer than a monomer and a template polymer does not occurs, while that of the complex between two of templates occurs infinitely fast. In this process, once a monomer forms the complex with a template, a substrate does not dissociate until complete the synthesis of the template. (This process requires a special design that specifically dissociates only the complex between the templates.)

We can reduce the process as,

$$X_{(1,m)} + X_{(L,s)} \to X_{(L,\tilde{s})} + X_{(L,s)},$$

where $\tilde{s}$ is the complemental sequence of the s. Then, the fraction of the template sequences x_i in the fixed point is consistent with one of the quasispecies model,

$$\dot{x}_i = \mathcal{N} \sum_j w_{ij} x_j - \phi x_i, \tag{4.22}$$

where $\mathcal{N}$ is a constant and $w_{ij} = r_j \binom{L}{h_{ij}} \mu^{L-h_{ij}} (1-\mu)^{h_{ij}}$. In this case, the accuracy of information (the fraction of the master sequences) declines with the length of the template same as the quasispecies model.

The error threshold for replicating template with the error correction

As is the previous discussion, due to the multi-step ligation, the error rate at the i-th bit of the template in the replication is modified as $\sim \mu^{L+1-i}$. In this case, the error threshold at which the growth rate of the master sequence is overwhelmed by that of the others is estimated by the condition,

$$\prod_{i=1}^{L} (1 - \mu^i) r_0 \sim r_1, \tag{4.23}$$

where r_0 and r_1 are the fitness of the master sequence and the other, respectively.

For infinite length L, $-\sum_{i=1}^{\infty} \log(1 - \mu^i)$ is bounded from above by $\sum_{i=1}^{\infty} i\mu^i$. From the condition that $W > \sum_{i=1}^{\infty} i\mu^i$, the critical error rate μ^* is at least

$$\mu^* > \frac{2 + 1/W - \sqrt{4/W + 1/W^2}}{2}, \tag{4.24}$$

where $W = \log(r_0/r_1)$. Therefore, even for the infinite length of template, the master sequence can persist under the finite error rate.

Chapter 5

Summary and Discussion

In this thesis, we studied the self-replicating system of biopolymers as a model for the origins of life. In the chapters 2-4, we studied the following issues;

- How long does it take for sufficiently long polymer to spontaneously emerge which can function as a catalyst? What is the condition for such emergence? (Chapter 2)
- What kind of polymer sequences will emerge? Under what condition the complex sequences which can code more information are selected? (Chapter 3)
- What is the relationship between the growth and the selection of template sequences? Under what condition can the information on the template attain robustness against errors and be sustained? (Chapter 4)

In such problems, kinetics of polymerization process have an essential role. We answered these questions in respect with properties of the environment, especially;

- Strength of nonequilibrium drive; i.e., the supply speed of monomers,
- Volume of the compartment; i.e., system size,

which provide parameters related to the kinetics.

Nonequilibrium drive

In chapter 3, we studied dependence of the selected sequence on the monomer supply rate in a model of template-directed ligation. Moreover, we also uncovered that the complexity of the selected polymer increases with the decrease of such flow rate. This dependency on the selection of complexity is due to the kinetic constraints in the reaction pathways to synthesize templates.

In chapter 4, on the other hand, we showed that the growth law and the consequence of the selection among the template polymer depend on the kinetic parameters of the polymerization process.The error correction mechanism akin to kinetic

proofreading prevents the error catastrophe during replication. The monomer supply and dissipation drive is necessary for such a correction mechanism to work. Similarly to kinetic proofreading mechanism, there is a trade-off between production speed and accuracy of information.

In both cases, the existence of the optimal flow rate or nonequilibrium drive exists for each mechanism to work efficiently for the self-replicating polymer system, although each has a different perspective.

System size

In chapter 2, we discussed transition time from the inactive state without a catalyst to the active state with catalyst, caused by stochastic noise due to the finiteness of system size. Then, we showed the existence of the optimal volume for the emergence of a self-replicating system with a long catalytic polymer. Also, we derived a formula that determines the optimal volume from the fixed point of the deterministic rate equation.

In the model in chapter 3, there also exists an optimal size for the selection of complex sequence to work, for the slow speed of flow. The transition in the model is due to the finiteness of the system size: if the size is too large, the transition to the attractor with the dominance of a complex sequence does not occur. On the other hand, if it is too small, the noise is so large that the system cannot stay at an attractor with the dominance of complex sequence.

The dependency on the system size for the model in chapter 4 is a future issue. As in section 1.2.1, for the evolution of exponentially growing sequence, the noise due to finiteness of system size is an important factor for the selection pressure. Also, for our template replication model with polymerization, the finiteness of the system size is expected to affect the robustness to error and the speed of evolution of the templates.

In this thesis, we showed that for both the system size and the strength of nonequilibrium drive, there is an optimal value for the self-replicating system of polymers to emerge. Such kinetic constraint for the prelife system will be important in the study of the origins of life, to examine the validity of various conflicting hypotheses for it.

The constant flow of monomer supply is realized in a laboratory condition by chemostat [42], where one can control the speed of such flow or the volume of the reactor easily. Of course, the parameters which determine the environment are not only the volume or speed of flow. For example, the temperature, pH or the metal ion concentration are considered to be important.

Since our model is abstract, it is difficult to quantitatively predict the optimal volume and speed of flow for realistic conditions. To estimate the actual value, experimental specification of the parameter values of biopolymers is necessary. On the contrary, elucidation on the properties of biopolymers required for a self-replicating

system to emerge and be sustained in a plausible environment should be essential to search for candidate biopolymers for the origins of life, among a vast variety of possibility.

Future issues

Energetics of polymerization

In this thesis, we studied the kinetics of polymerization rather than the energetics. It is necessary to discuss the energetics of polymer polymerization in considering a more realistic system.

As discussed in section 1.1, the exponential decay of the amount against the length of polymers is one of the problems for the emergence of the self-replicating system of biopolymers. Longer polymers are harder to synthesized. Therefore, how to supply energy from the external environment, other than the supply of activated monomers is a big problem in the origins of life. One of the possibilities is that energetic drive by the oscillation of the environmental condition. In fact in the experimental nonenzymatic synthesis of proteins and RNAs, oscillation of environmental conditions such as temperature oscillation or dry–wet cycles are considered to be essential. Then, it will be necessary to study theoretically what polymerization process can be accelerated by such oscillation whether the energy is converted efficiently into the chemical potential of the covalent bonds, and to search for the condition in which the synthesis of long polymers is possible. (For example, the minimum energy to require to proceed copolymerization process on the template is discussed (see section 1.3).)

Also, the extension of the model of the sequence selection discussed in chapter 3 to the model to include energetics is an important future issue. In the model in the section, the template polymers are assumed to be energetically homogeneous. In an actual polymer system, however, they are inhomogeneous. In preliminary study, the correlation between the speed of flow and the complexity of the selected sequence is suggested even in such case, but elucidation for the condition is important in future.

Emergence of genetic system

In this thesis, we discussed the first stages for the origins of life, replication of long polymers. However, the self-replicating system has to go through various other stages to evolve into the present life system. One of such stages is the acquirement of the genetic system, as discussed in section 1.2.3.

The kinetic origin of the molecule with the role of genetic information is discussed in Kaneko–Yomo model [88], In the model, a hypercycle with two components in a protocell is considered. If the symmetry of the replication rate between the two molecules is broken, the activity of minority molecule determines the growth rate of the whole cell. Thus the minority molecules plays a role for information carrier.

However, in the model, the growth rate of the total protocell itself is faster if there is no difference in the replication rates between the molecules. Therefore, how the symmetry in replication rates between such molecules is broken is not explained as a consequence of selection of a higher growth system.

The elementary process of the polymerization process as considered in this thesis could be a solution to the problem. As discussed in chapter 4, the replication rate of templates could be proportional to the powers of the concentration of itself. In that case, in contrast to the original model, the growth rate of the whole protocell is faster when the amounts of the two molecules are asymmetric. Therefore, considering the process of polymerization to synthesizing biomolecules, there is a possibility that a genetic system can be derived as a natural consequence of the chemical evolution of photocells.

Concluding remarks

In this thesis, we have considered a self-replication system of biopolymers in respect with the chemical reaction process of polymerization. We have elucidated some physical conditions for the emergence of self-replicating polymer systems. In the origins of life studies, such theoretical analysis is essential. By deriving stronger conditions for prebiotic life to emerge and sustain, we may elucidate history of life on the earth. Furthermore, uncovering constraints on the life will contribute to understand not only the origins of life but also the present life itself.

Acknowledgement

First of all, I would like to thank the supervising professor Kunihiko Kaneko. Not only he did gives me exciting ideas for research, but also patiently helped me to write papers.

During the long period of life as a graduate student in Kaneko Lab, I owe to many people. In particular, I appreciate Nobuto Takeuchi, Atsushi Kamimura, Tetsuhiro Hatakeyama, and Nen Saito. I also thank the members of the lab, Shunpei Yamamoto, Yuuma Fujimoto, Takuya Sato, Kenji Okubo, Naoto Nishiura, Jonathan Young, Takahiro Kosokabe, Masahiko Ueda, Hironobu Noguchi, and Yusuke Himeoka.

I also thank David Lacoste, Sergei Maslov, Paul Higgs, Nigel Goldenfeld, Dieter Braun, and Namiko Mitarai for fruitful discussion.

Finally, I thank my parents who have supported me.

Bibliography

[1] Pier Luigi Luisi. The emergence of life: from chemical origins to synthetic biology. Cambridge University Press, 2016. pages 3

[2] Tibor Gánti. The principles of life. Oxford University Press, 2003. pages 3

[3] L Hays, L Archenbach, J Bailey, R Barnes, J Barros, C Bertka, P Boston, E Boyd, M Cable, I Chen, et al. Nasa astrobiology strategy, 2015. pages 3

[4] John Von Neumann and Arthur Walter Burks. Theory of self-reproducing automata. University of Illinois Press Urbana, 1996. pages 3

[5] Robert L Forward. Dragon's egg. Del Rey, 2011. pages 3

[6] Greg Egan. Wang's carpets. New legends. New York: Tor Books, pp. 349–378, 1995. pages 3

[7] Charles Darwin. On the origin of species, 1859. Routledge, 1859. pages 3

[8] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. Nature, Vol. 171, No. 4356, pp. 737–738, 1953. pages 3

[9] Carl R Woese and George E Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences, Vol. 74, No. 11, pp. 5088–5090, 1977. pages 3

[10] Aleksandr Ivanovich Oparin. Origin of Life on the Earth. Oliver And Boyd; Edinburgh, 1957. pages 4

[11] Stanley L Miller, et al. A production of amino acids under possible primitive earth conditions. Science, Vol. 117, No. 3046, pp. 528–529, 1953. pages 4, 5, 25

[12] Walter Gilbert. The rna world. nature, Vol. 319, p. 618, 1986. pages 4

[13] Kelly Kruger, Paula J Grabowski, Arthur J Zaug, Julie Sands, Daniel E Gottschling, and Thomas R Cech. Self-splicing rna: autoexcision and autocyclization of the ribosomal rna intervening sequence of tetrahymena. cell, Vol. 31, No. 1, pp. 147–157, 1982. pages 4

[14] Julian Lewis Martin Raff Keith Roberts Bruce Alberts, Dennis Bray and James D Watson. Molecular biology of the cell, 1994. pages 4, 5

[15] Gerald F Joyce. The antiquity of rna-based evolution. Nature, Vol. 418, No. 6894, p. 214, 2002. pages 5

[16] Kepa Ruiz-Mirazo, Carlos Briones, and Andres de la Escosura. Prebiotic systems chemistry: new perspectives for the origins of life. Chemical Reviews, Vol. 114, No. 1, pp. 285–366, 2013. pages 5

[17] Falk Wachowius, James Attwater, and Philipp Holliger. Nucleic acids: function and potential for abiogenesis. Quarterly reviews of biophysics, Vol. 50, , 2017. pages 5

[18] Norio Kitadai and Shigenori Maruyama. Origins of building blocks of life: A review. Geoscience Frontiers, Vol. 9, No. 4, pp. 1117–1153, 2018. pages 5

[19] Gerald F Joyce and Jack W Szostak. Protocells and rna self-replication. Cold Spring Harbor perspectives in biology, Vol. 10, No. 9, p. a034801, 2018. pages 5

[20] Eric T Parker, Henderson J Cleaves, Jason P Dworkin, Daniel P Glavin, Michael Callahan, Andrew Aubrey, Antonio Lazcano, and Jeffrey L Bada. Primordial synthesis of amines and amino acids in a 1958 miller h2s-rich spark discharge experiment. Proceedings of the National Academy of Sciences, Vol. 108, No. 14, pp. 5526–5531, 2011. pages 5

[21] John Oró. Synthesis of adenine from ammonium cyanide. Biochemical and Biophysical Research Communications, Vol. 2, No. 6, pp. 407–412, 1960. pages 6

[22] Ronald Breslow. On the mechanism of the formose reaction. Tetrahedron Letters, Vol. 1, No. 21, pp. 22–26, 1959. pages 6

[23] Matthew W Powner, Béatrice Gerland, and John D Sutherland. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. Nature, Vol. 459, No. 7244, p. 239, 2009. pages 6

[24] R Bruce Martin. Free energies and equilibria of peptide bond hydrolysis and formation. Biopolymers: Original Research on Biomolecules, Vol. 45, No. 5, pp. 351–353, 1998. pages 6

[25] Norio Kitadai. Thermodynamic prediction of glycine polymerization as a function of temperature and ph consistent with experimentally obtained results. Journal of molecular evolution, Vol. 78, No. 3-4, pp. 171–187, 2014. pages 6

[26] Kirsten S Dickson, Christopher M Burns, and John P Richardson. Determination of the free-energy change for repair of a phosphodiester bond. Journal of Biological Chemistry, 2000. pages 6

[27] James P Ferris, Aubrey R Hill Jr, Rihe Liu, and Leslie E Orgel. Synthesis of long prebiotic oligomers on mineral surfaces. Nature, Vol. 381, No. 6577, p. 59, 1996. pages 6

[28] James P Ferris. Montmorillonite-catalysed formation of rna oligomers: the possible role of catalysis in the origins of life. Philosophical Transactions of the Royal Society of London B: Biological Sciences, Vol. 361, No. 1474, pp. 1777–1786, 2006. pages 6

[29] Paul J Flory. Principles of polymer chemistry. Cornell University Press, 1953. pages 6

[30] S Spiegelman, I Haruna, IB Holland, G Beaudreau, and Dk Mills. The synthesis of a self-propagating and infectious nucleic acid with a purified enzyme. Proceedings of the National Academy of Sciences, Vol. 54, No. 3, pp. 919–927, 1965. pages 6, 71

[31] DR Mills, RL Peterson, and Sol Spiegelman. An extracellular darwinian experiment with a self-duplicating nucleic acid molecule. Proceedings of the National Academy of Sciences of the United States of America, Vol. 58, No. 1, p. 217, 1967. pages 6, 71

[32] DL Kacian, DR Mills, FR Kramer, and S Spiegelman. A replicating rna molecule suitable for a detailed analysis of extracellular evolution and replication. Proceedings of the National Academy of Sciences, Vol. 69, No. 10, pp. 3038–3042, 1972. pages 6, 71

[33] Keqin Chen and Frances H Arnold. Enzyme engineering for nonaqueous solvents: random mutagenesis to enhance activity of subtilisin e in polar organic media. Nature Biotechnology, Vol. 9, No. 11, p. 1073, 1991. pages 6

[34] Tracey A Lincoln and Gerald F Joyce. Self-sustained replication of an rna enzyme. Science, Vol. 323, No. 5918, pp. 1229–1232, 2009. pages 7, 37

[35] Nilesh Vaidya, Michael L Manapat, Irene A Chen, Ramon Xulvi-Brunet, Eric J Hayden, and Niles Lehman. Spontaneous network formation among cooperative rna replicators. Nature, Vol. 491, No. 7422, p. 72, 2012. pages 7, 11, 37

[36] Wendy K Johnston, Peter J Unrau, Michael S Lawrence, Margaret E Glasner, and David P Bartel. Rna-catalyzed rna polymerization: accurate and general rna-templated primer extension. Science, Vol. 292, No. 5520, pp. 1319–1325, 2001. pages 7

[37] Aniela Wochner, James Attwater, Alan Coulson, and Philipp Holliger. Ribozyme-catalyzed transcription of an active ribozyme. science, Vol. 332, No. 6026, pp. 209–212, 2011. pages 7

[38] Hannes Mutschler, Aniela Wochner, and Philipp Holliger. Freeze–thaw cycles as drivers of complex ribozyme assembly. Nature chemistry, Vol. 7, No. 6, p. 502, 2015. pages 7

[39] David P Horning and Gerald F Joyce. Amplification of rna by an rna polymerase ribozyme. Proceedings of the National Academy of Sciences, Vol. 113, No. 35, pp. 9786–9791, 2016. pages 7

[40] Manfred Eigen. Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften, Vol. 58, No. 10, pp. 465–523, 1971. pages 7, 52

[41] Manfred Eigen and Peter Schuster. The hypercycle. Naturwissenschaften, Vol. 65, No. 1, pp. 7–41, 1978. pages 7, 26

[42] Aaron Novick and Leo Szilard. Description of the chemostat. Science, Vol. 112, No. 2920, pp. 715–716, 1950. pages 8, 38, 78

[43] Ira Leuthäusser. An exact correspondence between eigen ' s evolution model and a two-dimensional ising system. The Journal of chemical physics, Vol. 84, No. 3, pp. 1884–1885, 1986. pages 10

[44] Ira Leuthäusser. Statistical mechanics of eigen's evolution model. Journal of statistical physics, Vol. 48, No. 1-2, pp. 343–360, 1987. pages 10

[45] Pedro Tarazona. Error thresholds for molecular quasispecies as phase transitions: From simple landscapes to spin-glass models. Physical Review A, Vol. 45, No. 8, p. 6038, 1992. pages 10

[46] John Maynard Smith. Evolution and the Theory of Games. Cambridge university press, 1982. pages 11

[47] Eörs Szathmáry. On the propagation of a conceptual error concerning hypercycles and cooperation. Journal of Systems Chemistry, Vol. 4, No. 1, p. 1, 2013. pages 11

[48] Shoichi Toyabe and Dieter Braun. Templated ligation can create a hypercycle replication network. arXiv preprint arXiv:1802.06544, 2018. pages 12, 38

[49] Chikara Furusawa and Kunihiko Kaneko. Zipf ' s law in gene expression. Physical review letters, Vol. 90, No. 8, p. 088102, 2003. pages 12

[50] Günter von Kiedrowski. A self-replicating hexadeoxynucleotide. Angewandte Chemie International Edition in English, Vol. 25, No. 10, pp. 932–935, 1986. pages 12, 17, 37

[51] Eörs Szathmáry and Irina Gladkih. Sub-exponential growth and coexistence of non-enzymatically replicating templates. Journal of Theoretical Biology, Vol. 138, No. 1, pp. 55–58, 1989. pages 12

[52] Patrick Alfred Pierce Moran. Random processes in genetics. In Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 54, pp. 60–71. Cambridge University Press, 1958. pages 13

[53] Warren J Ewens. Mathematical population genetics 1: theoretical introduction, Vol. 27. Springer Science & Business Media, 2012. pages 13

[54] CW Gardiner. Stochastic methods. Springer-Verlag, Berlin–Heidelberg–New York–Tokyo, 1985. pages 13, 31, 56

[55] Nicolaas Godfried Van Kampen. Stochastic processes in physics and chemistry, Vol. 1. Elsevier, 1992. pages 13, 28, 31, 34, 56

[56] Yuichi Togashi and Kunihiko Kaneko. Transitions induced by the discreteness of molecules in a small autocatalytic system. Physical review letters, Vol. 86, No. 11, p. 2459, 2001. pages 14, 26

[57] Akinori Awazu and Kunihiko Kaneko. Discreteness-induced transition in catalytic reaction networks. Physical Review E, Vol. 76, No. 4, p. 041915, 2007. pages 14, 26

[58] Nen Saito and Kunihiko Kaneko. Theoretical analysis of discreteness-induced transition in autocatalytic reaction dynamics. Physical Review E, Vol. 91, No. 2, p. 022707, 2015. pages 14, 26, 35

[59] Akinori Awazu and Kunihiko Kaneko. Discreteness-induced slow relaxation in reversible catalytic reaction networks. Physical Review E, Vol. 81, No. 5, p. 051920, 2010. pages 14

[60] Andrew Duncan, Shuohao Liao, Tomáš Vejchodskỳ, Radek Erban, and Ramon Grima. Noise-induced multistability in chemical systems: Discrete versus continuum modeling. Physical Review E, Vol. 91, No. 4, p. 042111, 2015. pages 14, 35

[61] Tommaso Biancalani, Louise Dyson, and Alan J McKane. Noise-induced bistable states and their mean switching time in foraging colonies. Physical review letters, Vol. 112, No. 3, p. 038101, 2014. pages 14, 26, 35

[62] Farshid Jafarpour, Tommaso Biancalani, and Nigel Goldenfeld. Noise-induced mechanism for biological homochirality of early life self-replicators. Physical Review Letters, Vol. 115, No. 15, p. 158101, 2015. pages 14, 26, 53

[63] Martin Nowak and Peter Schuster. Error thresholds of replication in finite populations mutation frequencies and the onset of muller's ratchet. Journal of theoretical Biology, Vol. 137, No. 4, pp. 375–395, 1989. pages 14

[64] Akiko Kashiwagi, Itaru Urabe, Kunihiko Kaneko, and Tetsuya Yomo. Adaptive response of a gene network to environmental changes by fitness-induced attractor selection. PloS one, Vol. 1, No. 1, p. e49, 2006. pages 14, 53

[65] G Nicolis and I Prigogine. I (1977) self-organization in nonequilibrium systems. From Dissipative structures to Order through Fluctuations, Mir, Moscow, Russia. pages 15

[66] Freeman J Dyson. Origins of life. Cambridge University Press, 1985. pages 15, 26

[67] Stuart A Kauffman. Autocatalytic sets of proteins. Journal of theoretical biology, Vol. 119, No. 1, pp. 1–24, 1986. pages 16

[68] J Doyne Farmer, Stuart A Kauffman, and Norman H Packard. Autocatalytic replication of polymers. Physica D: Nonlinear Phenomena, Vol. 22, No. 1, pp. 50–67, 1986. pages 16, 26, 37

[69] Varun Giri and Sanjay Jain. The origin of large molecules in primordial autocatalytic reaction networks. PloS one, Vol. 7, No. 1, p. e29546, 2012. pages 17, 26, 37

[70] Elizaveta Guseva, Ronald N Zuckermann, and Ken A Dill. Foldamer hypothesis for the growth and sequence differentiation of prebiotic polymers. Proceedings of the National Academy of Sciences, p. 201620179, 2017. pages 17, 37

[71] Hue Sun Chan and Ken A Dill. ‘‘ sequence space soup ’’ of proteins and copolymers. The Journal of chemical physics, Vol. 95, No. 5, pp. 3775–3787, 1991. pages 17

[72] PHILIP W Anderson. Suggested model for prebiotic evolution: The use of chaos. Proceedings of the National Academy of Sciences, Vol. 80, No. 11, pp. 3386–3390, 1983. pages 17, 38

[73] DL Stein and PW Anderson. A model for the origin of biological catalysis. Proceedings of the National Academy of Sciences, Vol. 81, No. 6, pp. 1751–1753, 1984. pages 17

[74] Chrisantha Fernando, Günter Von Kiedrowski, and Eörs Szathmáry. A stochastic model of nonenzymatic nucleic acid replication: “ elongators ” sequester replicators. Journal of Molecular Evolution, Vol. 64, No. 5, pp. 572–585, 2007. pages 17, 38

[75] Harold Fellermann Shinpei Tanaka and Steen Rasmussen. Structure and selection in an autocatalytic binary polymer model. EPL (Europhysics Letters), Vol. 107, No. 2, p. 28004, 2014. pages 17

[76] Benedikt Obermayer, Hubert Krammer, Dieter Braun, and Ulrich Gerland. Emergence of Information Transmission in a Prebiotic RNA Reactor. Physical Review Letters, Vol. 107, No. 1, p. 018101, jun 2011. pages 17, 38

[77] Martin A Nowak and Hisashi Ohtsuki. Prevolutionary dynamics and the origin of evolution. Proceedings of the National Academy of Sciences, Vol. 105, No. 39, pp. 14924–14927, 2008. pages 17, 38

[78] Hisashi Ohtsuki and Martin Nowak. Prelife catalysts and replicators. Proceedings. Biological sciences / The Royal Society, Vol. 276, No. 1674, pp. 3783–90, nov 2009. pages 17, 38

[79] Alexei V Tkachenko and Sergei Maslov. Spontaneous emergence of autocatalytic information-coding polymers. The Journal of chemical physics, Vol. 143, No. 4, p. 045102, jul 2015. pages 17, 38

[80] Meng Wu and Paul G Higgs. Origin of self-replicating biopolymers: autocatalytic feedback can jump-start the rna world. Journal of molecular evolution, Vol. 69, No. 5, pp. 541–554, 2009. pages 17, 26

[81] George Francis Gause. The struggle for existence. Courier Corporation, 2003. pages 18

[82] András Szilágyi, István Zachar, and Eörs Szathmáry. Gause's principle and the effect of resource partitioning on the dynamical coexistence of replicating templates. PLoS computational biology, Vol. 9, No. 8, p. e1003193, 2013. pages 18

[83] Eric R Pianka. On r-and k-selection. The American Naturalist, Vol. 104, No. 940, pp. 592–597, 1970. pages 18

[84] Atsushi Kamimura and Kunihiko Kaneko. Negative scaling relationship between molecular diversity and resource abundances. Physical Review E, Vol. 93, No. 6, p. 062419, 2016. pages 18, 53

[85] Eörs Szathmáry and László Demeter. Group selection of early replicators and the origin of life. Journal of theoretical biology, Vol. 128, No. 4, pp. 463–486, 1987. pages 19

[86] Eörs Szathmáry and John Maynard Smith. The major evolutionary transitions. Nature, Vol. 374, No. 6519, pp. 227–232, 1995. pages 19

[87] John Maynard Smith and Eors Szathmary. The major transitions in evolution. Oxford University Press, 1997. pages 20

[88] Kunihiko Kaneko and Tetsuya Yomo. On a kinetic origin of heredity: minority control in a replicating system with mutually catalytic molecules. Journal of theoretical biology, Vol. 214, No. 4, pp. 563–576, 2002. pages 20, 53, 79

[89] Pierre Gaspard. Template-directed copolymerization, random walks along disordered tracks, and fractals. Physical review letters, Vol. 117, No. 23, p. 238101, 2016. pages 21

[90] David Andrieux and Pierre Gaspard. Nonequilibrium generation of information in copolymerization processes. Proceedings of the National Academy of Sciences, Vol. 105, No. 28, pp. 9516–9521, 2008. pages 21

[91] Michael Nguyen and Suriyanarayanan Vaikuntanathan. Design principles for nonequilibrium self-assembly. Proceedings of the National Academy of Sciences, Vol. 113, No. 50, pp. 14231–14236, 2016. pages 21

[92] John J Hopfield. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. Proceedings of the National Academy of Sciences, Vol. 71, No. 10, pp. 4135–4139, 1974. pages 22, 73

[93] Jacques Ninio. Kinetic amplification of enzyme discrimination. Biochimie, Vol. 57, No. 5, pp. 587–595, 1975. pages 22, 73

[94] Charles H Bennett. Dissipation-error tradeoff in proofreading. BioSystems, Vol. 11, No. 2-3, pp. 85–91, 1979. pages 22, 73

[95] Pablo Sartori and Simone Pigolotti. Kinetic versus energetic discrimination in biological copying. Physical review letters, Vol. 110, No. 18, p. 188101, 2013. pages 22, 73

[96] Arvind Murugan, David A Huse, and Stanislas Leibler. Speed, dissipation, and error in kinetic proofreading. Proceedings of the National Academy of Sciences, Vol. 109, No. 30, pp. 12034–12039, 2012. pages 22, 73

[97] Pablo Sartori and Simone Pigolotti. Thermodynamics of error correction. Physical Review X, Vol. 5, No. 4, p. 041039, 2015. pages 22, 73

[98] Harvey F Lodish, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, David Baltimore, James Darnell, et al. Molecular cell biology, Vol. 4. Citeseer, 2000. pages 25

[99] Yoshihiro Furukawa, Hiromoto Nakazawa, Toshimori Sekine, Takamichi Kobayashi, and Takeshi Kakegawa. Nucleobase and amino acid formation through impacts of meteorites on the early ocean. Earth and Planetary Science Letters, Vol. 429, pp. 216–222, 2015. pages 25

[100] Richard Wolfenden and Mark J Snider. The depth of chemical time and the power of enzymes as catalysts. Accounts of chemical research, Vol. 34, No. 12, pp. 938–945, 2001. pages 25

[101] Maarten C Boerlijst and Paulien Hogeweg. Spiral wave structure in prebiotic evolution: hypercycles stable against parasites. Physica D: Nonlinear Phenomena, Vol. 48, No. 1, pp. 17–28, 1991. pages 26

[102] Sanjay Jain and Sandeep Krishna. Autocatalytic sets and the growth of complexity in an evolutionary model. Physical Review Letters, Vol. 81, No. 25, pp. 5684–5687, 1998. pages 26

[103] Daniel Segré, Dafna Ben-Eli, and Doron Lancet. Compositional genomes: prebiotic information transfer in mutually catalytic noncovalent assemblies. Proceedings of the National Academy of Sciences, Vol. 97, No. 8, pp. 4112–4117, 2000. pages 26

[104] Kunihiko Kaneko. On recursive production and evolvabilty of cells: Catalytic reaction network approach. Advanced in Chemical Physics, Vol. 130, pp. 543–598, 2005. pages 26, 37

[105] Jun Ohkubo, Nadav Shnerb, and David A. Kessler. Transition phenomena induced by internal noise and quasi-absorbing state. Journal of the Physical Society of Japan, Vol. 77, No. 4, p. 044002, 2008. pages 26

[106] Thierry Dauxois, Francesca Di Patti, Duccio Fanelli, and Alan J McKane. Enhanced stochastic oscillations in autocatalytic reactions. Physical Review E, Vol. 79, No. 3, p. 036112, 2009. pages 26

[107] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. The journal of physical chemistry, Vol. 81, No. 25, pp. 2340–2361, 1977. pages 29, 40

[108] R Graham and T Tél. On the weak-noise limit of fokker-planck models. Journal of statistical physics, Vol. 35, No. 5, pp. 729–748, 1984. pages 32

[109] Daniel T Gillespie. The chemical langevin equation. The Journal of Chemical Physics, Vol. 113, No. 1, pp. 297–306, 2000. pages 34

[110] Pier Luigi Luisi. The emergence of life: from chemical origins to synthetic biology. Cambridge University Press, 2006. pages 35

[111] Steen Rasmussen. Protocells. Mit Press, 2009. pages 35

[112] Erwin Schrodinger. What is life? University Press: Cambridge, 1943. pages 37

[113] Hubert Krammer, Friederike M Möller, and Dieter Braun. Thermal, autonomous replicator made from transfer rna. Physical review letters, Vol. 108, No. 23, p. 238104, 2012. pages 37

[114] Jan W Sadownik, Elio Mattia, Piotr Nowak, and Sijbren Otto. Diversification of self-replicating molecules. Nature chemistry, Vol. 8, No. 3, pp. 264–269, 2016. pages 37, 38

[115] James M Carothers, Stephanie C Oestreich, Jonathan H Davis, and Jack W Szostak. Informational complexity and functional activity of rna structures. Journal of the American Chemical Society, Vol. 126, No. 16, pp. 5130–5137, 2004. pages 37

[116] Julien Derr, Michael L Manapat, Sudha Rajamani, Kevin Leu, Ramon Xulvi-Brunet, Isaac Joseph, Martin A Nowak, and Irene A Chen. Prebiotically plausible mechanisms increase compositional diversity of nucleic acid sequences. Nucleic acids research, Vol. 40, No. 10, pp. 4711–22, may 2012. pages 37, 38

[117] Manfred Eigen and Peter Schuster. A principle of natural self-organization. Naturwissenschaften, Vol. 64, No. 11, pp. 541–565, 1977. pages 37

[118] Péter Szabó, István Scheuring, Tamás Czárán, and Eörs Szathmáry. In silico simulations reveal that replicators with limited dispersal evolve towards higher efficiency and fidelity. Nature, Vol. 420, No. 6913, p. 340, 2002. pages 37

[119] Yoshiya J Matsubara and Kunihiko Kaneko. Optimal size for emergence of self-replicating polymer system. Physical review. E, Vol. 93, No. 3-1, p. 032503, mar 2016. pages 37, 53

[120] Grant Kinsler, Sam Sinai, Nicholas Keone Lee, and Martin A Nowak. Prebiotic selection for motifs in a model of template-free elongation of polymers within compartments. PloS one, Vol. 12, No. 7, p. e0180208, 2017. pages 37, 38

[121] D. L. Stein and P. W. Anderson. A model for the origin of biological catalysis. Proceedings of the National Academy of Sciences of the United States of America, Vol. 81, No. March, pp. 1751–1753, 1984. pages 38

[122] Shinpei Tanaka, Harold Fellermann, and Steen Rasmussen. Structure and selection in an autocatalytic binary polymer model. EPL (Europhysics Letters), Vol. 107, No. 2, p. 28004, jul 2014. pages 38

[123] Christine He, Isaac Gállego, Brandon Laughlin, Martha A Grover, and Nicholas V Hud. A viscous solvent enables information transfer from gene-length nucleic acids in a model prebiotic replication cycle. Nature Chemistry, 2016. pages 38

[124] Alexei V Tkachenko and Sergei Maslov. Onset of natural selection in autocatalytic heteropolymers. arXiv preprint arXiv:1710.06385, 2017. pages 48, 51

[125] Chikara Furusawa and Kunihiko Kaneko. A generic mechanism for adaptive growth rate regulation. PLoS Computational Biology, Vol. 4, No. 1, p. e3, 2008. pages 53

[126] Edward N Trifonov. Making sense of the human genome. Structure and methods: proceedings of the Sixth Conversation in the Discipline Biomolecular Stereodynamics held at the State University of New York at Albany, June 6-10, 1989/edited by RH Sarma & MH Sarma, 1990. pages 53

[127] O Popov, DM Segal, and Edward N Trifonov. Linguistic complexity of protein sequences as compared to texts of human languages. Biosystems, Vol. 38, No. 1, pp. 65–74, 1996. pages 53

[128] Yuri L Orlov and Vladimir N Potapov. Complexity: an internet resource for analysis of dna sequence complexity. Nucleic acids research, Vol. 32, No. suppl_2, pp. W628–W633, 2004. pages 53

[129] Claude Elwood Shannon. A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review, Vol. 5, No. 1, pp. 3–55, 2001. pages 53

[130] Andrei N Kolmogorov. On tables of random numbers. Sankhyā: The Indian Journal of Statistics, Series A, pp. 369–376, 1963. pages 53

[131] Seth Lloyd and Heinz Pagels. Complexity as thermodynamic depth. Annals of physics, Vol. 188, No. 1, pp. 186–213, 1988. pages 53

[132] Andrei Gabrielian and Alexander Bolshoy. Sequence complexity and dna curvature. Computers & chemistry, Vol. 23, No. 3-4, pp. 263–274, 1999. pages 53

[133] Philipp Baaske, Franz M Weinert, Stefan Duhr, Kono H Lemke, Michael J Russell, and Dieter Braun. Extreme accumulation of nucleotides in simulated hydrothermal pore systems. Proceedings of the National Academy of Sciences, Vol. 104, No. 22, pp. 9346–9351, 2007. pages 73

[134] Moritz Kreysing, Lorenz Keil, Simon Lanzmich, and Dieter Braun. Heat flux across an open pore enables the continuous replication and selection of oligonucleotides towards increasing length. Nature chemistry, Vol. 7, No. 3, p. 203, 2015. pages 73