Sequence motif dynamics in RNA pools

Johannes Harth-Kitzerow,[1,2,3,4] Tobias Göppel,[2,4] Ludwig Burger,[2,4] Torsten A. Enßlin,[1,3,4] and Ulrich Gerland[2,4]

[1] *Max Planck Institute for Astrophysics Garching, Karl-Schwarzschild-Str. 1, 85741 Garching, Germany;*

[2] *Department of Bioscience, TUM School of Natural Sciences, Technical University Munich, 85748 Garching, Germany;*

[3] *Ludwig-Maximilians-Universität München, Fakultät für Physik, Geschwister-Scholl-Platz 1, 80539 München LMU, Germany;*

[4] *Excellence Cluster Origins, Boltzmannstraße 2, 85748 Garching, Germany*

ABSTRACT

In RNA world scenarios, pools of RNA oligomers form strongly interacting, dynamic systems, which enable molecular evolution. In such pools, RNA oligomers hybridize and dehybridize, ligate and break, ultimately generating longer RNA molecules, which may fold into catalytically active ribozymes. A key process for the elongation of RNA oligomers is templated ligation, which can occur when two RNA strands are adjacently hybridized onto a template strand. Detailed simulations of the dynamics in RNA pools involve a large variety of possible sequences and reactions. Here, we develop a reduced description of these complex dynamics within the space of sequence motifs. We then explore to what extent our reduced description can capture the behavior of detailed simulations that account for the full dynamics in the space of RNA strands. Towards this end, we project the dynamics into a motif space, which accounts only for the abundance of all possible four-nucleotide motifs. A system of ordinary differential equations describes the dynamics of those motifs. Its control parameters are effective rate constants for reactions in motif space, which we obtain from the rate constants for the processes underlying the full dynamics in the space of RNA strands. We find that these reduced motif space dynamics indeed capture important aspects of the informational dynamics of RNA pools in sequence space. This approach could also provide a framework to rationalize and interpret features of the sequence dynamics observed in experimental systems.

INTRODUCTION

Pools of inter-reacting RNA molecules are a possible starting point for molecular evolution and play a key role in the RNA world hypothesis for the origin of life [2, 8, 16, 17, 20]. The collective behavior of RNA pools, i. e., mixtures of RNA molecules in solution, arises from a complex interplay of intra- and inter-molecular processes. Non-covalent interactions, such as the formation of base-pairs and the stacking of base-pairs into helices, govern the self-folding of RNA strands and the binding between different RNA strands ('hybridization' of strands). At the same time, chemical reactions involving covalent bonds govern processes such as template-directed ligation and polymerization, as well as cleavage of RNA strands via hydrolysis. As a consequence, RNA pools are systems with an enormous range of relevant timescales, and a tremendous combinatorial complexity. It is currently neither feasible to track the detailed dynamics of RNA pools experimentally, nor to accurately predict these dynamics computationally. However, the outcome of these dynamics can be assayed experimentally, for instance with mass spectrometry [11] or sequencing approaches [1]. Computational approaches and theoretical analyses can then help to interpret such experiments, and to extrapolate into experimentally inaccessible parameter regimes or to intractable timescales.

In principle, the dynamics of RNA pools could be modeled at various levels of description. However, a very detailed molecular description is computationally intractable when the long-time behavior of a sizable pool is of interest. A hybrid approach like OxRNA [27], which simulates the dynamics of a coarse-grained nucleotide-level RNA model by combining kinetic Monte Carlo and Molecular Dynamics techniques, would be well suited to describe, e. g., the assembly of the RNA complexes required for template-directed ligation. However, this level of description is still too detailed for the slow timescales associated, in particular, with the ligation and cleavage reactions. Prior computational and theoretical work [3, 4, 7, 9, 12, 15, 19, 21–26] has therefore used more coarse-grained descriptions that do not explicitly keep track of the polymer conformation of the RNA (or DNA) molecules in the system. Instead, the descriptions for instance keep track of the sequence composition and the strand length distribution of the pool, and estimate the rates for the different kinetic processes included in the model using sequence-dependent base-pairing free energies. For instance, template-directed ligation can occur only when two strands are adjacently hybridized to a template

strand. Hence the kinetic rate for this process is proportional to the number of such RNA complexes, which depends on their thermodynamic stability.

While the dynamics of RNA pools is rooted in the polymer physics and chemistry of RNA strands, it can also be regarded from a purely information-centric perspective. For instance, we can consider a string of nucleotides with a specific sequence and count how often it occurs at any position on the RNA strands within a pool. How does the distribution of such sequence motifs evolve in an RNA pool, given a set of intra- and inter-molecular processes? And how do these sequence motif dynamics depend on the rate constants for the different processes? Furthermore, is it possible to describe important features of the sequence motif dynamics in RNA pools with a closed set of dynamical equations, without explicitly describing the underlying physico-chemical processes acting on the RNA strands? These questions motivate the analysis we present here.

As a concrete example and proof of principle, we base our analysis on the 'RNA reactor' model and simulation framework [7, 9, 19], which models the interplay of four processes: RNA hybridization, de-hybridization, cleavage, and template-directed ligation. Focusing on the case of 4-nucleotide motifs, we devise motif rate equations that describe the time evolution of motif concentrations under the simultaneous action of the four physico-chemical processes of the RNA reactor. Our approach can be regarded as a generalization of the 2-nucleotide motif dynamics of Ref. [22]. Importantly, for our 4-nucleotide motif dynamics, we can establish a direct relation between its kinetic parameters and those of the underlying physico-chemical processes acting on the RNA strands. Four is the minimal motif length, for which this is possible, since the RNA reactor model includes a kinetic effect in ligation that depends not only on the identity of the nucleotides at the ligation site, but also on the neighboring nucleotides to both sides. This kinetic stalling of template-directed ligation after mismatches has been established experimentally [10, 18] and can have a significant impact even on the qualitative behavior of RNA pools [7].

The mapping between the kinetic parameters of the full RNA reactor model and those of the motif rate equations permits us to test which aspects of the dynamics are retained by the reduced motif description, and which are lost. Towards this end, we consider each of the five different scenarios of Ref. [7] for the RNA reactor also within the motif rate equations. In the following Methods section, we therefore first briefly summarize the RNA reactor model before developing the reduced description in motif space. The comparison between the two

levels of description then follows in the Results section.

METHODS

In this section, we introduce motif rate equations as chemical rate equations for motif concentrations, and determine their parameters from those of the strand-based 'RNA reactor' model [7, 9, 19], see Fig. 1 for illustration. The 'RNA reactor' model describes the kinetics of oligonucleotide strands that can take on different configurations in linear (non-branched) complexes, but do not fold onto themselves. These kinetics are simulated using the Gillespie algorithm [5, 6], with kinetic rate constants that are chosen consistent with the thermodynamics of hybridization between the oligonucleotide strands (in the following, $\Delta\mathcal{G}_{\mathrm{tot}}$ denotes the total free energy difference between a configuration where a single strand is bound in a complex and the configuration where this strand is unbound). Here, we specifically use the energy model of Ref. [7] for binary nucleotide alphabets, which allows for an energy bias for alternating sequences, such that fully hybridized blocks of two by two nucleotides have a hybridization energy difference $\Delta\gamma$ if they consist of perfectly matching alternating 2-nucleotide motifs compared to perfectly matching homogeneous 2-nucleotide motifs. (If the configuration involves a dangling end, i. e., a 2-nucleotide motif is hybridized to the end of a strand such that one nucleotide has no binding partner, the corresponding energy bias for alternating sequences is denoted by δ_ϵ.) Furthermore, single strands can be cleaved at any bond with a rate constant k_{cut}, independent of the nucleotide sequence context, and templated ligation is characterized by a ligation rate constant k_{lig}. In the presence of mismatches (non-complementary hybridization of nucleotides) directly at the ligation site, the ligation rate constant is reduced by a stalling factor σ_1, while next-nearest neighbor mismatches stall ligation with the factor σ_2. We consider a closed system without influx and outflux of strands or nucleotides.

Motif concentrations

Before we introduce the motif rate equations, we specify how we track the motif concentrations, which we group into a motif concentration vector $\vec{c}$. Towards that end, we define the motifs we track, introducing conventions to prevent double counting of motifs,

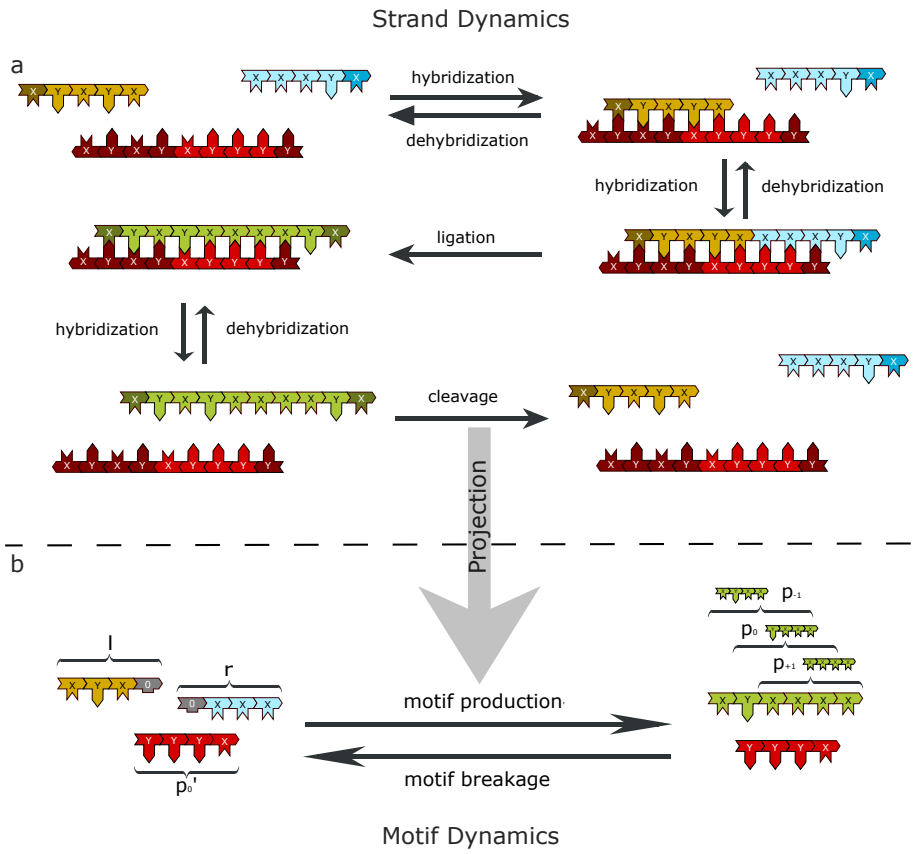

FIG. 1. Kinetic processes of the strand-based 'RNA reactor' model [7, 9, 19] and their projection onto motif kinetics. We consider heteropolymer strands consisting of the two complementary nucleotides X and Y, which could represent RNA or DNA nucleotides. (a) Illustration of the reactions within the strand reactor, which include the hybridization, dehybridization, templated ligation, and cleavage of oligonucleotide strands. Different colors represent different strands (not covalently linked) and four-nucleotide motifs are highlighted by different shades. (b) On the level of sequence motifs, the kinetics are based on motif production and motif breakage reactions. In the four-nucleotide motifs, the ends of strands are represented by 0. Templated ligation produces longer strands containing more motifs, thus contributing to motif production. Conversely, cleavage of strands contributes to motif breakage. Hybridization is not explicitly tracked within the motif description, but implicitly included in the rate constants for motif production and breakage by assuming a hybridization-dehybridization equilibrium and deducing the corresponding dissociation constants.

and present how we convert strand numbers into motif concentrations.

To establish a proof of principle for our approach, we use a minimal binary nucleotide alphabet $\mathcal{A} = \{\mathrm{X}, \mathrm{Y}\}$, consisting of the two generic complementary nucleotides X and Y. These could represent, e.g., the RNA nucleotides A and U, or the DNA nucleotides A and T. Based on this alphabet, we consider sequence motifs that are directed (chemically from the 5' to the 3' end of a strand). To track beginnings and ends of strands, we introduce the additional symbol 0, and denote with $\mathcal{A}_0 = \mathcal{A} \cup \{0\}$ the alphabet including 0. Since the ligation rate constant in the strand reactor simulation depends on the two nucleotides to the left and the two nucleotides to the right of the ligation site, the minimal motif length required to project the strand kinetics on the motif space is four. We thus choose to focus on the 4-nucleotide motif dynamics. We track monomers and dimers explicitly, while all longer strands are tracked by their containing 4-nucleotide motifs, including their beginning and end motifs, i. e. motifs that start, or respectively end, with an empty spot symbol 0, but contain only letters elsewhere (Fig. 1b). We refer to beginning motifs simply as beginnings, end motifs as ends, and call 4-nucleotide motifs that do not contain a 0 but only letters continuations.

We avoid double counting of motifs by introducing the convention that the second spot of a 4-nucleotide motif is always a letter. The first and the last spot can be a letter or empty. This allows for ends, beginnings, and continuations, as well as for monomers and dimers. The third spot is only empty for monomers such that empty spots do not occur in between letters.

To determine motif concentrations from a given state of the strand reactor, we define the concentration of a single particle, $c_\Delta = \frac{1}{V_\mathrm{R}}$, i. e., the concentration of one complex in the strand reactor with a volume V_R. The motif concentrations are then obtained by multiplying c_Δ with the counts of all 4-nucleotide motifs in the strands of the reactor (following the above conventions). Note that our description of the motif dynamics in terms of motif rate equations (to be introduced below) corresponds to a mean field approach, which implicitly assumes the thermodynamic limit of the reactor volume V_R going to infinity at constant concentrations, such that the dynamics can be described by chemical rate equations instead of a chemical master equation (see also Appendix C 1).

Motif rate equations

We now turn to the construction of the motif rate equations for a closed reactor, without influx or outflux. We consider the limit, in which ligation is much slower than hybridization and dehybridization, such that a hybridization-dehybridization equilibrium is reached. The fraction of single-stranded versus hybridized strands is then determined by dissociation constants,

$$K_{\mathrm{D}}(C) := e^{\beta \Delta \mathcal{G}_{\mathrm{tot}}(C)}, \tag{1}$$

for each complex configuration C. In this limit, we incorporate templated ligation together with the dissociation constants in the motif production terms, $\vec{f}_{\mathrm{lig}}$. Analogously, we derive motif breakage terms, $\vec{f}_{\mathrm{break}}$, from the dissociation constants and strand cleavage. Thus, the total motif rate equation contains these two classes of reaction terms:

$$\partial_\tau \vec{c} = \vec{f}_{\mathrm{lig}}(\vec{c}) + \vec{f}_{\mathrm{break}}(\vec{c}). \tag{2}$$

Motif production, the first set of terms in the motif rate equation captures hybridization and dehybridization implicitly and ligation explicitly. The corresponding motif production rate constants are denoted $\tilde{k}_{\mathrm{ext}}(p_{\mathrm{c}}, t)$, with produced motif p_{c} onto the templating motif t. It describes the process of two connecting end strands, one ending reactant coming with an open 3' end, $l = (l_1, l_2, l_3, 0)$ (l stands for left strand, defined by $l_1 \in \mathcal{A}_0, l_2, l_3 \in \mathcal{A}$), respectively $l = (0, l_2, 0, 0)$ in case it is a monomer ($l_1, l_3 = 0$), and one beginning reactant with an open 5' end, $r = (0, r_2, r_3, r_4)$ (r stands for right strand, defined by $r_2 \in \mathcal{A}, r_3, r_4 \in \mathcal{A}_0$) that hybridize adjacent to each other onto a template strand with motif $t = (t_1, t_2, t_3, t_4)$, see Figure 1. Then they ligate to the products

$$p_{\mathrm{l}} := p_{\mathrm{l}}(l, r) := (l_1, l_2, l_3, r_2), \tag{3}$$

$$p_{\mathrm{c}} := p_{\mathrm{c}}(l, r) := (l_2, l_3, r_2, r_3), \text{ and} \tag{4}$$

$$p_{\mathrm{r}} := p_{\mathrm{r}}(l, r) := (l_3, r_2, r_3, r_4) \tag{5}$$

and eventually dehybridize again. In case the left (right) reactant is a monomer, there is no first produced motif p_{l} (third produced motif p_{r}).

Since we only consider nearest neighbour contributions to the effective ligation rate, $\tilde{k}_{\mathrm{ext}}(p_{\mathrm{c}}, t)$ does neither depend on the first nucleotide of the ending strand (l_1) nor the last

nucleotide of the leaving strand (r_4). The reaction takes place with the total motif production rate $\Lambda_{l,r,t}$,

$$\Lambda_{l,r,t} = \tilde{k}_{\mathrm{ext}}(p_{\mathrm{c}}(l,r),t)\ c_l\ c_r\ c_t. \tag{6}$$

Details of its calculation are in Appendix A 1. With this, the motif production terms split up into five sets of terms. Two for the ending and beginning reactant, $\vec{f}_{\mathrm{lig_l}}, \vec{f}_{\mathrm{lig_r}}$ and three terms for the three produced motifs, collected in $\vec{f}_{\mathrm{lig_p}}$.

$$\vec{f}_{\mathrm{lig}}(\vec{c}) = \vec{f}_{\mathrm{lig_l}}(\vec{c}) + \vec{f}_{\mathrm{lig_r}}(\vec{c}) + \vec{f}_{\mathrm{lig_p}}(\vec{c}). \tag{7}$$

Both, the ending as well as the leaving reactant, are consumed in this reaction. This leads to a negative contribution in the motif rate equation,

$$\left[\vec{f}_{\mathrm{lig_l}}(\vec{c})\right]_{l'} = -\sum_{r',t} \Lambda_{l',r',t} \tag{8}$$

$$\left[\vec{f}_{\mathrm{lig_r}}(\vec{c})\right]_{r'} = -\sum_{l',t} \Lambda_{l',r',t} \tag{9}$$

for ends, i. e., possible ending reactants including monomers, and beginnings, i. e., possible beginning reactants. For all other motifs, $\vec{f}_{\mathrm{lig_l}}$ and $\vec{f}_{\mathrm{lig_r}}$ become zero.

For the three produced motifs $(p_{\mathrm{l}}, p_{\mathrm{c}}, p_{\mathrm{r}})$, we get positive contributions in the motif rate equation,

$$\left[\vec{f}_{\mathrm{lig_p}}(\vec{c})\right]_{p''} = \sum_{p_4,p_5,t} \Lambda_{l,r,t}\delta_{p'',p_{\mathrm{l}}(l,r)} + \sum_{p_0,p_5,t} \Lambda_{l,r,t}\delta_{p'',p_{\mathrm{c}}(l,r)} + \sum_{p_0,p_1,t} \Lambda_{l,r,t}\delta_{p'',p_{\mathrm{r}}(l,r)}, \tag{10}$$

without a contribution of any other motif. For convenience, we chose a compact notation without stating the summation ranges explicitly. The ranges following from our convention are discussed in detail in Appendix A 1.

By splitting the motif production terms into five subterms for the two reactants and three motifs, we have set up the motif production terms of the motif rate equation. For the motif production rate constants, we use the extension rate constants derived in [7] and discussed in Appendix A 2 for complexes up to length four.

Breakage leads to two terms in the motif rate equation. First, a negative term $\vec{f}^{\,-}_{\mathrm{break}}$ for the motif that breaks and, second, a positive term $\vec{f}^{\,+}_{\mathrm{break}}$ for motifs that result from breakage. Before we discuss the terms for the motif rate equation, we draw the connection to the strand reactor simulation to derive the breakage rates from cleavage rate constant and

dissociation constant. Breakage of motifs is the reaction that models the effect of cleavage in the strand reactor simulation. For the motif rate equations, we have to adjust the rate constant by taking the possible protection due to hybridization into account. In the strand reactor simulation, cleavage is modeled by a sequence independent cleavage rate constant k_{cut} for any single stranded segment. Since we do not store hybridizations in the motif dynamics, we need to reconstruct their occupation using the dissociation constant which gives the fraction of free versus occupied strands as described above.

In principle, there are three breakage spots in every 4-nucleotide motif. Since we lose correlations longer than the motif length, we approximate the stability of a hybridization site by the dissociation constant for a single 4-nucleotide motif interpreted as a tetramer, i. e. assuming it to be a strand of four nucleotides only, as we did for the motif production rate constants. In fact, hybridization can be stronger, if the overlap is actually longer, which consequently results in an underestimation of the number of double stranded segments in the motif rate equations. We obtain the breakage rate by multiplying the breakage rate constant with the dissociation constant. Since the fraction of double stranded segments is underestimated, the breakage rate constant is thus overestimated in the motif rate equatitons. We leave the introduction of a correction term for this for future work.

Now that we have clarified the connection to cleavage, we can calculate the breakage rate constants. In a 4-nucleotide motif there are three breakage spots that we denote by the index $i \in \{1, 2, 3\}$. For the central breakage spot ($i = 2$), we use the template-averaged dissociation constant $\tilde{K}_{\text{D}}(m)$, i. e.the dissociation constant of the strand $m = (m_1, m_2, m_3, m_4)$ for all configurations in which the central bond (m_2, m_3) is hybridized to any other motif. It is calculated analogously to the sequence-averaged dissociation constant in [7]: – the inverse average of inverse dissociation constants of the single configurations $C = [m, m']$ – but by averaging over the template for a given sequence. In case of the template-averaged dissociation constants those are the hybridizations onto different templates m'.

$$\left(\tilde{K}_{\text{D}}(m)\right)^{-1} = \sum_{m'} \left(K_{\text{D}}([m, m'])\right)^{-1} \tag{11}$$

Then the effective cleavage rate constant $\tilde{k}_{\text{cut}}$ is given by the cleavage rate constant k_{cut} multiplied by the template-averaged dissociation constant.

$$\tilde{k}_{\text{cut}}(m) = k_{\text{cut}} \tilde{K}_{\text{D}}(m) \tag{12}$$

We approximate the (central) motif breakage rate constant $\tilde{k}^{(2)}_{\text{cut}}(m)$ by the effective cleavage rate constant ($\tilde{k}_{\text{cut}}(m)$). A central breakage also implies left and right breakage. The breakage rate constants of the left and the right bond need to be determined from the central breakage consistently. For that we continue the left (right) breaking motif to the left (right) with the left (right) continuation frequency ϕ_- (ϕ_+), which is given by the fraction of the concentration of the specific continuation divided by the sum of concentrations of all other possible continuations including a beginning (end),

$$\phi_-(m = (m_0, m_1, m_2, m_3)) := \frac{c_{(m_0,m_1,m_2,m_3)}}{\sum_{m_0' \in \mathcal{A}_0} c_{(m_0',m_1,m_2,m_3)}}, \tag{13}$$

$$\phi_+(m = (m_2, m_3, m_4, m_5)) := \frac{c_{(m_2,m_3,m_4,m_5)}}{\sum_{m_5' \in \mathcal{A}_0} c_{(m_2,m_3,m_4,m_5')}}. \tag{14}$$

Then we marginalize over the continuations to the left (right) and corresponding breakage rate constants to get effective breakage rate constants $\tilde{k}^{(i)}_{\text{cut}}(m)$ for the left ($i = 1$), central ($i = 2$) and right ($i = 3$) bond [28],

$$\tilde{k}^{(2)}_{\text{cut}}(m) = \tilde{k}_{\text{cut}}(m), \tag{15}$$

$$\tilde{k}^{(1)}_{\text{cut}}(m) = \sum_{\substack{m_0 \in \mathcal{A}_0 \\ (m_1,\ldots,m_4)=m}} \tilde{k}^{(2)}_{\text{cut}}((m_0, m_1, m_2, m_3))\phi_-((m_0, m_1, m_2, m_3)), \tag{16}$$

$$\tilde{k}^{(3)}_{\text{cut}}(m) = \sum_{\substack{m_5 \in \mathcal{A}_0 \\ (m_1,\ldots,m_4)=m}} \tilde{k}^{(2)}_{\text{cut}}((m_2, m_3, m_4, m_5))\phi_+((m_2, m_3, m_4, m_5)). \tag{17}$$

Given the effective breakage rate constants, we can now set up the rates for the breaking motifs $\vec{f}^{\,-}_{\text{break}}$ and the resulting ones $\vec{f}^{\,+}_{\text{break}}$

$$\vec{f}^{\,-}_{\text{break}}(\vec{c})_m = -\sum_{i=1}^{\ell-1} \Gamma^{(i)}_{(m)}, \text{ with} \tag{18}$$

$$\Gamma^{(i)}_{(m)} := \tilde{k}^{(i)}_{\text{cut}}(m) c_m, \tag{19}$$

and

$$\vec{f}^{\,+}_{\text{break}}(\vec{c})_{m=(0,m_2,m_3,m_4)} = \sum_{x \in \mathcal{A}} \Gamma^{(1)}_{((x,m_2,m_3,m_4))}, \tag{20}$$

$$\vec{f}^{\,+}_{\text{break}}(\vec{c})_{m=(0,m_2,0,0)} = \sum_{x_0 \in \mathcal{A}_0, x_1 \in \mathcal{A}} \left(\Gamma^{(2)}_{((x_0,x_1,m_2,0))} + \Gamma^{(2)}_{((0,m_2,x_1,x_0))} \right), \tag{21}$$

$$\vec{f}^{\,+}_{\text{break}}(\vec{c})_{m=(m_1,m_2,m_3,0)} = \sum_{x \in \mathcal{A}} \Gamma^{(3)}_{((m_1,m_2,m_3,x))} \tag{22}$$

with $m_2, m_3 \in \mathcal{A}$ and $m_1, m_4 \in \mathcal{A}_0$. The central breakage term (21) only produces motifs in case they are monomers, as all other broken motifs are already considered in the left and right breakage terms.

In summary, the total motif rate equation is given by the sum of motif production terms and breakage terms. Hybridization and dehybridization are captured in all rates implicitly by approximating hybridization-dehybridization equilibrium via dissociation constants. Approximating motifs by strands of the same length as the motifs, we calculated corresponding motif production rate constants and breakage rate constants from the parameters of the strand simulation. For details on its implementation, see Appendix A 3.

RESULTS AND DISCUSSION

To test and validate our projection of the strand dynamics onto motif rate equations, we compare the respective dynamics for five different parameter sets. We directly compare the motif concentration trajectories, but also use observables such as motif entropy and mean strand length, where the latter can be reconstructed from the motif concentrations as we will show. Before discussing the results, we describe the five different parameter sets and choice of initial conditions.

Parameter Sets and Initial Conditions

In order to compare the sequence motif dynamics in the strand reactor simulation to the solutions of the motif rate equations, we use five parameter sets labeled 0 to 4, which focus on different combinations of three features described above [7]: ligation stalling (s), cleavage (c), and hybridization bias (h). For each feature, we use two parameter values, which either essentially deactivate or activate the corresponding feature, e.g., ($\neg$s, s) for ligation stalling, see Table I. Parameter set 0 deactivates all three features ($\neg\mathrm{s}, \neg\mathrm{c}, \neg\mathrm{h}$). Parameter set 1 has ligation stalling activated ($\mathrm{s}, \neg\mathrm{c}, \neg\mathrm{h}$), while parameter set 2 additionally activates cleavage ($\mathrm{s}, \mathrm{c}, \neg\mathrm{h}$). In parameter set 3, cleavage is deactivated again, but hybridization bias is activated ($\mathrm{s}, \neg\mathrm{c}, \mathrm{h}$), and finally in parameter set 4, all three features are activated ($\mathrm{s}, \mathrm{c}, \mathrm{h}$).

In most cases, we use the same initial conditions as in Ref. [7]: The closed RNA reactor contains a total mass of 5000 nucleotides, initially predominantly as monomers (2460 of

Parameter		Value 0	Value 1
Ligation Stalling	σ_1	1.0	0.05
	σ_2	1.0	1.0
Cleavage	k_{cut}	$4.24 \cdot 10^{-18}$	$9.056 \cdot 10^{-11}$
Hybridization Bias	$\Delta\gamma$	0	-0.3
	δ_ϵ	0	-0.15

Parameter Set	Ligation Stalling	Cleavage	Hybridization Bias
0	0	0	0
1	1	0	0
2	1	1	0
3	1	0	1
4	1	1	1

TABLE I. Parameter sets used for the strand reactor simulations. The top table shows, for each varied parameter, the two values referred to as '0' and '1' below. The bottom table lists the chosen value combinations for the five parameter sets.

each nucleotide, X and Y) in an unbiased mix with a few dimers (10 dimers of each kind, XX, XY, YX, YY). The reactor volume is such that the concentration of a single molecule is $c_\Delta = 2 \cdot 10^{-6} c_0$, where $c_0 = 1 \frac{\text{mol}}{\text{L}}$ denotes the reference concentration. In the case of parameter set 2, we also probe the dependence of the dynamics on the initial conditions, and thus vary the initial conditions away from these default values. For the motif rate equations, we convert the initial conditions into the corresponding motif concentrations, as described above, i.e., we scan each strand for all its containing 4-nucleotide motifs and multiply the resulting total motif counts with c_Δ.

Templated Ligation without Cleavage, Ligation Stalling and Hybridization Bias

We first consider the case of negligible cleavage, without ligation stalling nor hybridization bias (parameter set 0). Fig. 2 illustrates the resulting motif dynamics by showing motif concentration trajectories for X monomers, YY dimers, XYX beginnings, and XXXX motifs (to capture the behavior on many scales, both the concentration and the time axis have

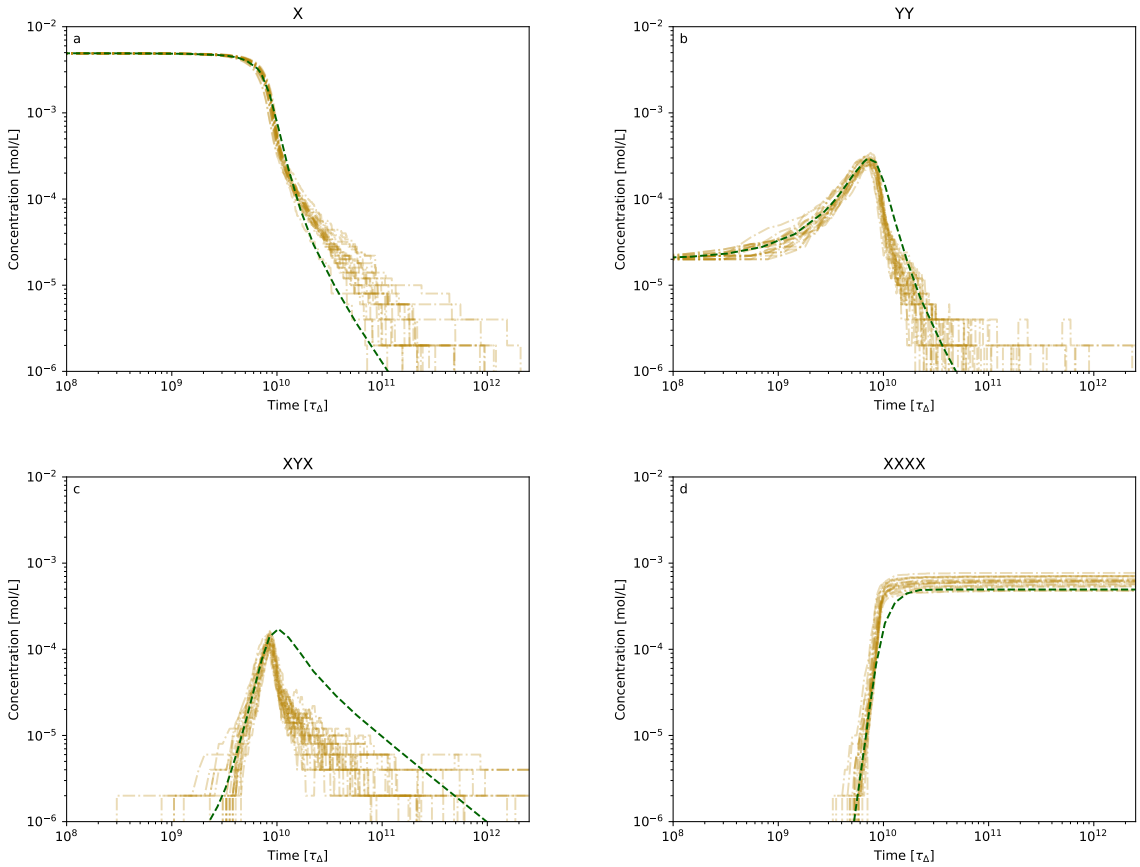

FIG. 2. Motif concentrations in the strand reactor simulation (gold) and motif rate equations (green) of the motifs as stated for parameter set 0, i. e.without hybridization bias, without ligation stalling and negligible cleavage.

a logarithmic scale). Here, the solution of the motif rate equations (green dashed lines) is superimposed on an ensemble of stochastic trajectories from the strand reactor simulation.

The trajectories in Fig. 2 reflect the elongation dynamics of strands via templated ligation. Monomers, which are initially dominant in the pool are depleted, generating dimers and longer strands. The dimer concentrations first rise from their small initial values, then reach a peak, before dimers are depleted together with the monomers. The 3-nucleotide beginnings (trajectory for XYX) display a concentration peak shortly after the dimers, with a concurrent rise of the 4-nucleotide motifs, which then reach a stationary value. The elongation dynamics is directly seen in Fig. 3, which shows the mean strand length $\overline{L}$ as a function of time. Here, the mean-field approximation to the mean strand length (green dashed line) is computed from the motif rate equations as described in Appendix C 3.

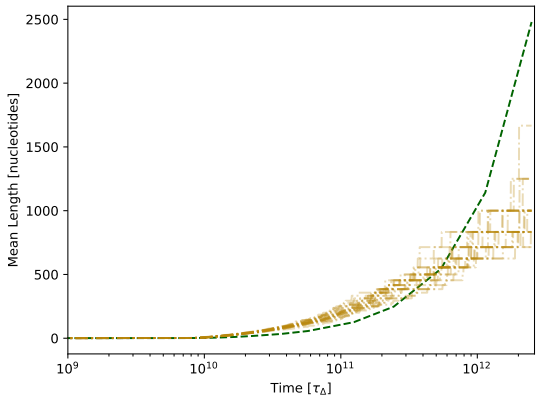

FIG. 3. Mean length of the strand reactor simulation (dash dotted gold, mean solid gold) and the motif rate equations (dashed green) for parameter set 0 without hybridization bias, without ligation stalling and negligible cleavage.

As seen from Figs. 2 and 3, the motif rate equations capture many aspects of the stochastic strand dynamics in the RNA reactor for parameter set 0. In the initial phase, when the mean strand length is below the motif length, the agreement is quantitative. Subsequently, the elongation dynamics within the motif rate equations lag behind the strand reactor simulation. The origin of this lag lies in the fact that our motif rate equations cannot account for the slow dehybridization of strands longer than the motif length of four nucleotides: The motif rate equations effectively overestimate the dissociation constants, and thus underestimate the extension rate constants, for longer strands. On the other hand, when long strands end up in unproductive double-stranded configurations, they cannot serve as templates anymore. The interplay between these effects has been investigated in detail in Ref. [19]. The elongation behavior within the motif rate equations indeed catches up with respect to the strand reactor simulation at large times. This is also due to the fact that the motif rate equations describe the thermodynamic limit of infinite particles, such that strands in the motif rate equations can grow infinitely long, while in the strand reactor simulation, growth is limited by the total mass of 5000 nucleotides in the finite RNA reactor (see Fig. 3). In principle, the finite size of the RNA pool could also affect the dynamics in the strand reactor already at the onset of growth. However, in our case the pool is large enough to avoid such an effect, as estimated in Appendix C 2.

Besides the elongation dynamics, the dynamics of the RNA reactor in sequence space

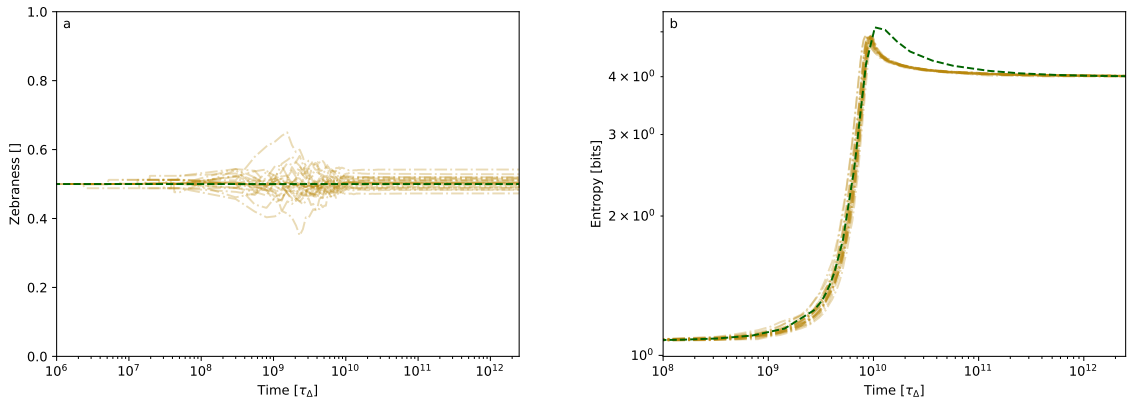

FIG. 4. System-level zebraness (a) and motif entropy (b) of the strand reactor simulation (dash dotted gold, mean solid gold) and the motif rate equations (dashed green) for parameter set 0 qualitatively agree.

is of key interest to characterize its behavior. Towards this end, Ref. [7] introduced the "zebraness" Z as a system-level observable measuring an asymmetry that can spontaneously arise in an RNA reactor. The system-level zebraness is defined as the fraction of alternating 2-mers, 'XY' and 'YX' within the sequence pool. Hence Z takes values between 0 and 1. If the pool contains only homogeneous motifs ('XX', 'XX'), the zebraness is 0, while it is 1, if the pool contains only zebra-like motifs ('XY', 'YX'). If the pool is unbiased and all motifs occur with the same frequency, the zebraness is 0.5. In terms of 4-nucleotide motifs, the system-level zebraness is calculated via

$$Z = \frac{\sum_{\substack{x_1,x_4\in\mathcal{A}_0,\\ m_2,m_3\in\mathcal{A}}} c_{x_1,m_2,m_3,x_4}(1-\delta_{m_2,m_3})}{\sum_{\substack{x_1,x_4\in\mathcal{A}_0,\\ m_2,m_3\in\mathcal{A}}} c_{x_1,m_2,m_3,x_4}} \,. \tag{23}$$

Another relevant observable for the distribution in sequence space is the motif entropy. The dynamics of both observables, the zebraness Z and the motif entropy, is shown in Fig. 4 for parameter set 0. Since in this case there is no hybridization energy bias for alternating motifs, the system-level zebraness Z stays at its initial level of 0.5 for the motif rate equations(Fig. 4a, dashed green line). This is also the ensemble average of the strand reactor simulation trajectories, which experience some stochastic fluctuations at the onset of growth, at which ligations of longer strands become more frequent than ligations of two monomers.

In contrast, the motif entropy shows a distinctive feature in its dynamics (Fig. 4b),

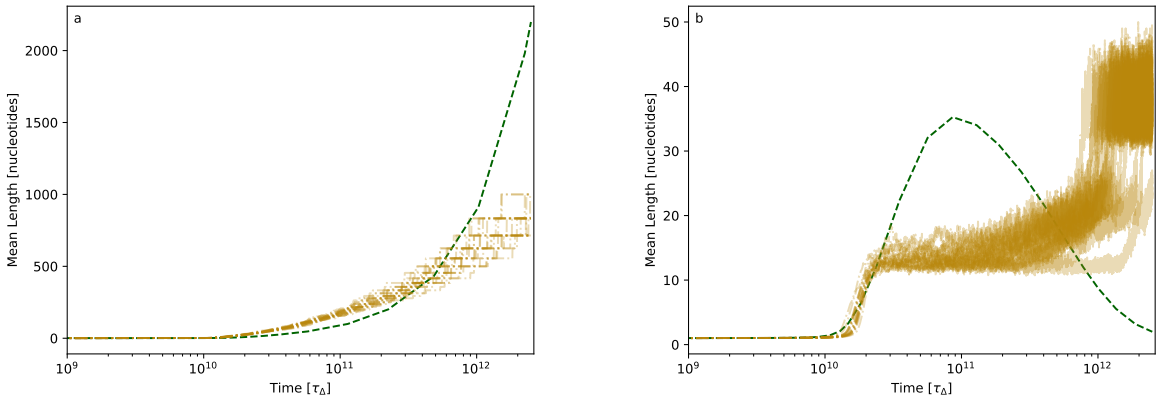

FIG. 5. Time evolution of the mean strand length, as in Fig. 3, but for (a) parameter set 1 and (b) parameter set 2, where ligation stalling is activated, and in (b) also cleavage is activated.

recapitulating the behavior already observed above for the individual motifs (Fig. 2). After a growth phase, the entropy displays a transient maximum, and then relaxes to a stationary value. The motif rate equations quantitatively describe the growth phase, lag behind the strand reactor simulation during the relaxation phase, but ultimately reach the correct stationary value.

Taken together, most aspects of the strand dynamics are captured well by the motif rate equations for parameter set 0. Expected deviations due to stronger hybridization are present, but do not affect the qualitative behavior of the dynamics and are also quantitatively small. Obviously, solving the motif rate equations amounts to a much lower computational effort than the strand reactor simulation (by about four orders of magnitude for the present case). We next explore whether these conclusions remain true also in the presence of cleavage, ligation stalling, and hybridization bias. Since we found that for the sake of comparing the motif rate equations to the strand reactor simulation, the motif entropy recapitulates the behavior of the individual motif concentration trajectories, we will focus on the motif entropy as an aggregate observable for the remaining parameter sets.

Ligation Stalling, Cleavage and Hybridization Bias

We first consider the effect of ligation stalling (parameter set 1), which slows down templated ligation in the presence of mismatches close to the ligation site. Accordingly, the

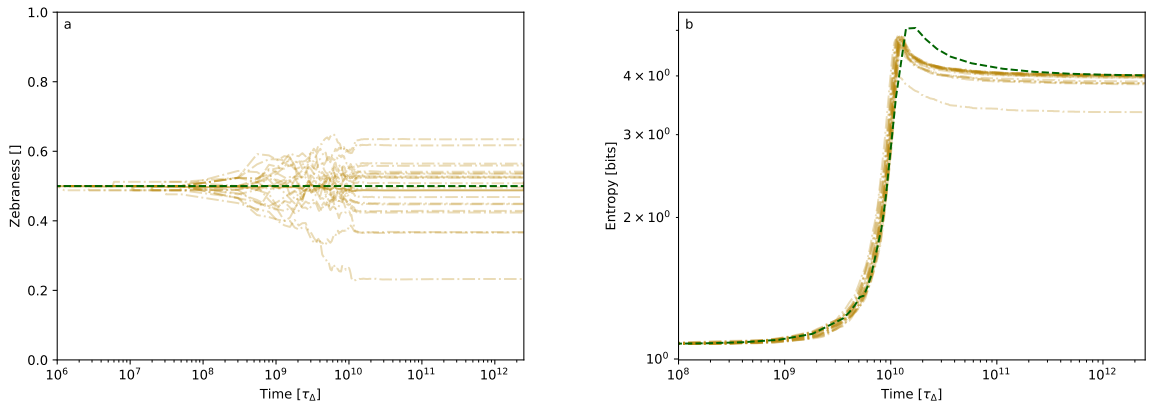

FIG. 6. Time evolution of (a) the system-level zebraness and (b) the motif entropy, as in Fig. 4, but for parameter set 1, in which ligation stalling is activated.

time evolution of the mean strand length, shown in Fig. 5a, is slightly slower compared to the case without ligation stalling (Fig. 3), but otherwise not significantly altered. The same holds true for the time evolution of the motif entropy (Fig. 6b) compared to the case without ligation stalling (Fig. 4b). For both observables, the motif rate equations capture the behavior of the strand reactor simulation to the same degree as for parameter set 0.

A clear difference in the behavior is seen only in the system-level zebraness (Fig. 6a). As in the case without ligation stalling (Fig. 4a), the strand reactor simulation produces fluctuations during the growth phase, but the resulting bias in the zebraness then persists in the presence of ligation stalling. Ref. [7] revealed that this is due to the tendency of ligation stalling to suppress ligation in the presence of mismatches, which would otherwise counteract the bias. Hence the persisting bias in the individual strand reactor simulation trajectories corresponds to an effect of locked-in fluctuations. Fluctuations are not accounted for in the mean field description of the motif rate equations, such that their dynamics consistently stay at a zebraness of 0.5 in Fig. 6a.

Next, we turn to the effect of cleavage by considering parameter set 2. Cleavage drastically changes the time evolution of the mean strand length in the strand reactor simulation (Fig. 5b), creating a plateau shortly after the onset of growth, followed by a second phase of growth settling at a higher plateau at much later times. The origin of this behavior, studied in detail in Ref. [7], is visible in the time evolution of the zebraness and entropy shown in Fig. 7: The combination of ligation stalling and cleavage leads to spontaneous symmetry

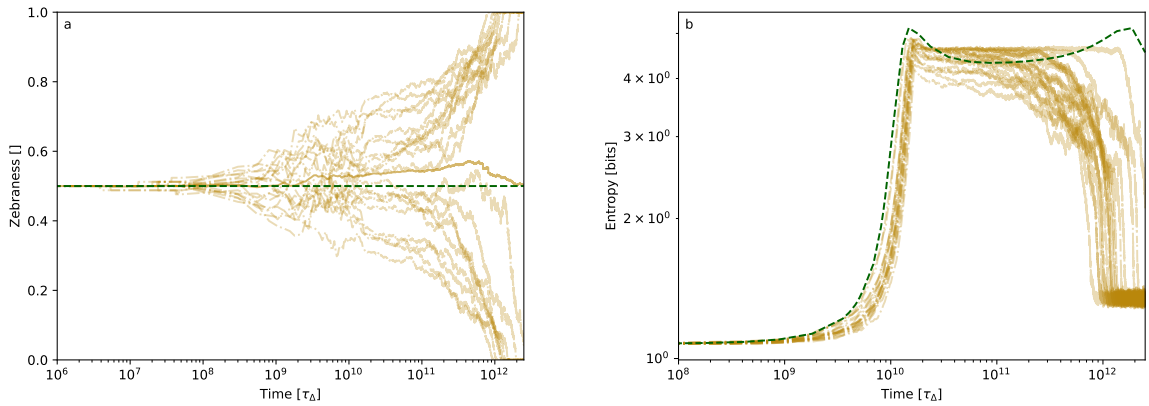

FIG. 7. Time evolution of (a) the system-level zebraness and (b) the motif entropy, as in Fig. 4, but for parameter set 2, in which ligation stalling and cleavage are activated.

breaking in sequence space, where the sequence composition of the pool ultimately converges either to purely homogeneous sequences (zebraness 0) or purely alternating sequences (zebraness 1), accompanied by a drop in the sequence entropy (Fig. 7b). Both cases lead to the formation of more stable double-strands, which are less susceptible to cleavage, producing longer strands on average (Fig. 5b).

This intricate interplay between stochastic effects and hybridization-inhibited cleavage cannot be captured by the motif rate equations. It comes as no surprise then that there are significant qualitative deviations between motif rate equations and strand reactor simulation in Figs. 5b and 7. The origin of these deviations lies in the two inherent limitations of the motif rate equations already discussed above, (i) the missing effects of finite pool size and stochasticity within the mean field approach, and (ii) the missing kinetic effects caused by longer hybridization sites. Nevertheless, while strands are still short, up to a mean strand length of about 12 nucleotides, the motif rate equations describe the strand reactor simulation. The zebraness always remains at the unbiased value of 0.5 within the strand reactor simulation, consistent with the mean field approach that cannot account for spontaneous symmetry breaking in a finite pool. As a result, the entropy increases within the motif rate equations, when it decreases within the strand reactor simulation, since the strand reactor simulation trajectories decide for one zebraness regime leading to a decrease of entropy, while the motif rate equations solutions take both states at the same time leading to an increase of entropy. The difference in the behavior of the mean length at large times

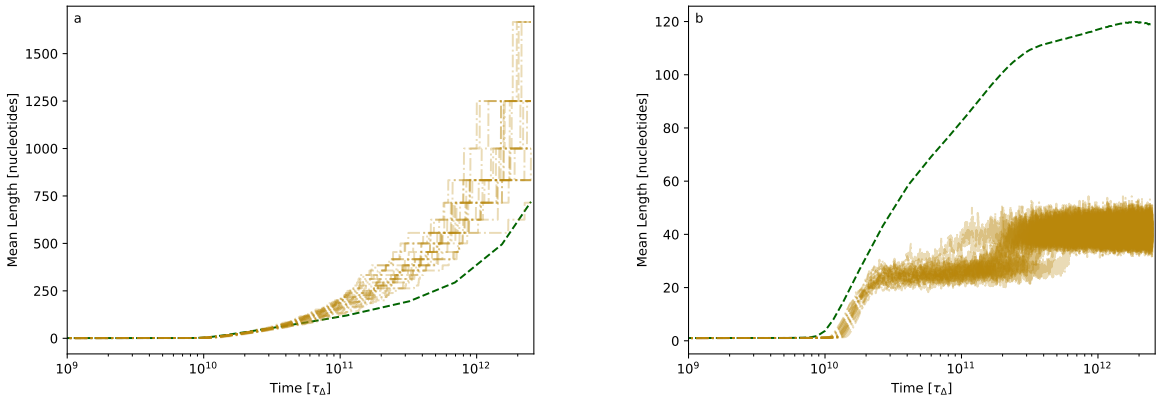

FIG. 8. Time evolution of the mean strand length, as in Fig. 5, but for (a) parameter set 3 and (b) parameter set 4, probing the additional effect of a hybridization bias.

(Fig. 5b) is explained by the missing protective effect of stable long hybridizations against cleavage within the motif rate equations.

So far, we did not obtain an emerging zebraness bias within the motif rate equations, since they cannot capture the spontaneous symmetry breaking that occurs within the strand reactor simulation in parameter set 2. In principle, symmetry breaking in sequence space could emerge within the motif rate equations in two ways. First, via the energy parameters, specifically by introducing a hybridization bias as we will do next. Second, by amplification of a small bias in the initial conditions, which we investigate further below.

In both, parameter sets 3 and 4, the hybridization bias is activated, which means that the hybridization of alternating sequences is energetically more favorable, and hence more stable, than for homogeneous sequences. The hybridization bias does not have a strong effect on the time evolution of the mean strand length within the strand reactor simulation: For deactivated cleavage (parameter set 3), the hybridization bias enhances growth only slightly (Fig. 8a) compared with the case without hybridization bias (Fig. 5a). In the presence of cleavage (parameter set 4), the hybridization bias does not alter the two-step growth to a mean length of around 40 nucleotides observed already without hybridization bias (Fig. 5b vs. Fig. 8b).

In contrast, the predicted dynamics of the mean length within the motif rate equations is substantially changed by the hybridization bias. Whereas they overestimate strand growth without cleavage in the absence of the hybridization bias (Fig. 5a), they underestimate

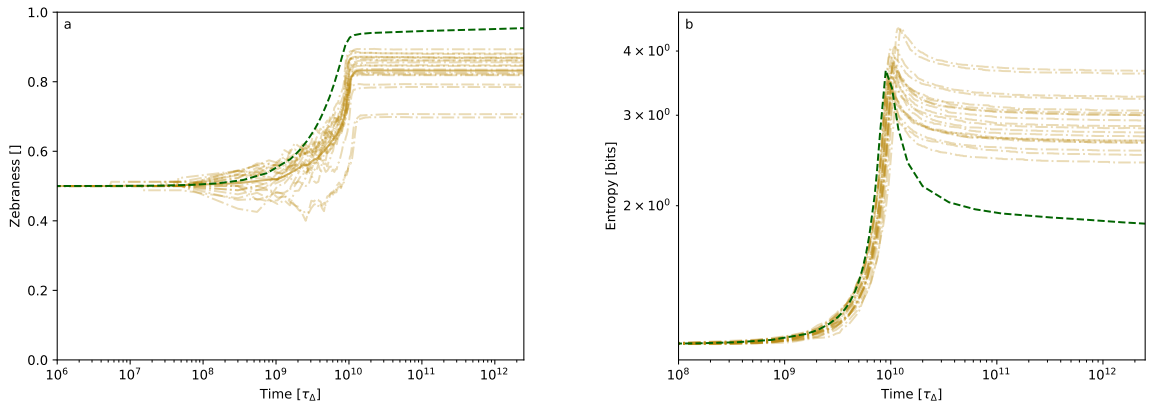

FIG. 9. Time evolution of (a) the system-level zebraness and (b) the motif entropy, as in Fig. 6, but for parameter set 3, in which the hybridization bias is activated.

growth in the presence of the hybridization bias (Fig. 8a). With cleavage, the motif rate equations are similarly sensitive to the hybridization bias (Fig. 5b vs. Fig. 8b). This illustrates that although the estimation of the mean strand length from the motif rate equations (Appendix C 3) is quite accurate at short lengths, it is generally not robust. This comes as no surprise, since the motif rate equations focus on the informational dynamics on the scale of short motifs.

Indeed, the motif rate equations capture these informational dynamics semi-quantitatively for both parameter sets, 3 and 4, see Fig. 9 and Fig. 10, respectively. The hybridization bias leads to an emerging zebraness bias for both cases, without cleavage (Fig. 9a) and with cleavage (Fig. 10a). In the latter case, the zebraness bias even reaches its maximum value of $Z = 1$ in the long time limit, corresponding to a pool consisting only of alternating sequences. This behavior is also captured within the motif rate equations. Without cleavage, the motif rate equations reproduce the dynamics of the strand reactor simulation relatively well, but exaggerate the zebraness bias in the long time limit. Since high or low zebraness corresponds to lower motif entropy than the mixed state, the motif entropy also decreases further for the motif rate equations than for the strand reactor simulation (Fig. 9b). This deviation can be rationalized by the fact that the motif rate equations cannot account for the slow dehybridization of strands longer than the motif length of four nucleotides, since the loss of binding energy by mismatches can be compensated by the hybridization of longer strands, allowing more strands with mismatches to be extended by ligation within the strand

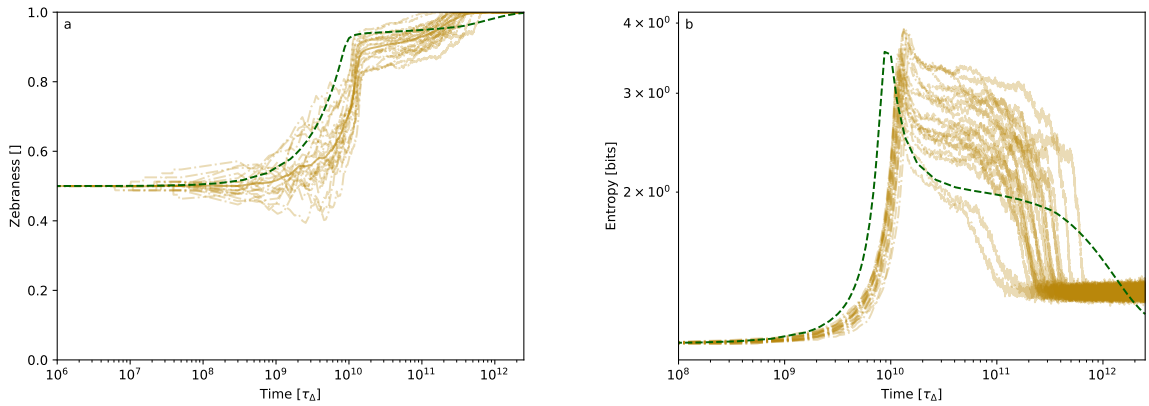

FIG. 10. Time evolution of (a) the system-level zebraness and (b) the motif entropy, as in Fig. 7, but for parameter set 4, in which the hybridization bias is activated.

reactor simulation, which increases the motif entropy.

Taken together, although the motif rate equations cannot capture effects due to finite size or stochasticity, they can otherwise describe the informational dynamics of the strand reactor simulation in motif space on a semi-quantitative level. Given that the description via motif rate equations is computationally much more efficient than strand reactor simulation, this suggests that motif rate equations are a useful tool to rapidly explore the parameter dependence of the informational dynamics. This approach could be used, for instance, to identify regions within the parameter space that potentially exhibit interesting behaviors, which can then be studied in detail using strand reactor simulation. As a small illustration for this approach, we briefly consider the dependence on the initial conditions in the following.

Probing fluctuation effects via biased initial conditions

We saw above that the spontaneous, noise-driven symmetry breaking in sequence space, which occurs within the strand reactor simulation for parameter set 2, cannot be described by the motif rate equations (Fig. 7). However, we can still use the deterministic description of the motif rate equations to probe whether an initial fluctuation that breaks the symmetry will grow or be attenuated, or whether the bias remains at the initial level.

Towards this end, we probe the behavior of the motif rate equations for different zebraness biases in the initial sequence pool. As before, the initial pool is dominated by monomers,

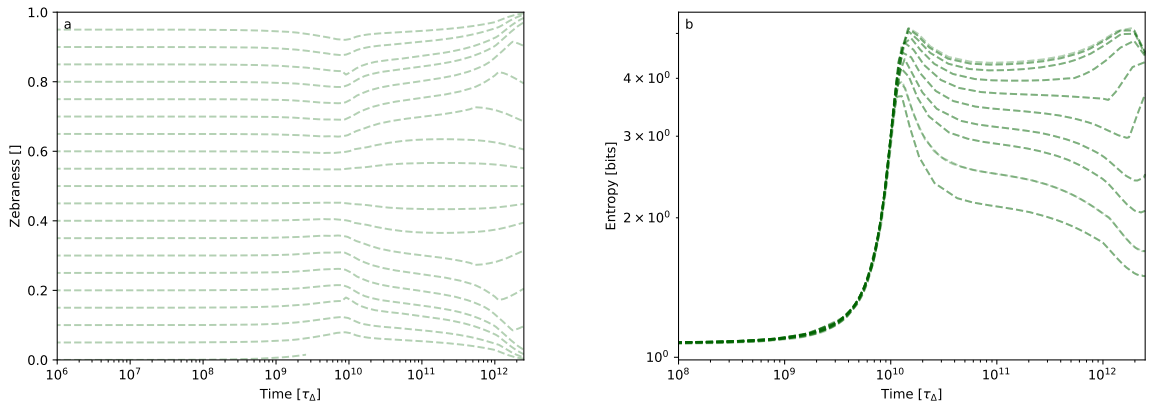

FIG. 11. Time evolution of (a) the system-level zebraness and (b) the motif entropy within the motif rate equations for parameter set 2, starting from different initial conditions, in which the dimers within the initial pool have different biases towards alternating or homogeneous sequences.

with an equal number of X and Y. However, instead of the unbiased initial condition with 10 dimers of each type, we start, e.g., with 18 alternating dimers (9 each of 'XY' and 'YX') and 22 homogeneous dimers (11 each of 'XX' and 'YY'). Fig. 11 shows the resulting informational dynamics obtained from the motif rate equations with parameter set 2. The plots also show the curves for any other possible combination of alternating and homogeneous dimers in the initial pool (at the same total of 40 dimers).

As seen from the zebraness and motif entropy in Fig. 11, the sequence pool tends to more biased states after the initial growth. For strong initial biases, mimicking a large fluctuation amplitude in the stochastic system, the zebraness even approaches almost 0 or 1 in the long-time limit. This confirms that the spontaneous symmetry breaking in the stochastic strand reactor simulation has a signature in the deterministic dynamics of the motif rate equations. Such signatures can be used for rapid parameter screens based on the motif rate equations.

CONCLUSION AND OUTLOOK

We analyzed the informational dynamics in pools of polynucleotides, which bind to each other and react via ligation and cleavage processes. Focusing on the level of short sequence motifs, we developed a system of coupled motif rate equations to directly describe the motif

dynamics extracted from full stochastic simulations on the level of polynucleotide strands, using the previously established 'RNA reactor' simulation framework [7, 9, 19].

Our comparison of the predictions of the motif rate equations with the strand reactor simulation was based on five different parameter sets with distinct behaviors, and on four observables: motif concentrations, motif entropy, 'zebraness' [7], and mean strand length. As might be expected based on the projection of the dynamics into motif space, the motif rate equations can generally not accurately predict the time evolution of the mean strand length in the strand reactor simulation. In contrast, we found that many aspects of the informational dynamics were well described by the motif rate equations. This is noteworthy, in particular, because the motif rate equations do not involve any fit parameters. Instead, all parameters of the motif rate equations are directly determined from the parameters of the RNA reactor model. We also found systematic deviations between the motif dynamics predicted by the motif rate equations and the strand reactor simulation. These could be attributed to effects depending on the influence of longer hybridization sites than the motif length, as well as stochasticity and finite system size effects.

Since the numerical solution of the motif rate equations is computationally much more affordable than the strand reactor simulation, and the motif rate equations capture many aspects of the informational dynamics on a semi-quantitative level, it will be useful to scan the parameter space for phenomena of interest, which can then be studied in more detail within the strand reactor simulation. This approach will have to be fully developed in the future. However, as a first illustration, we showed that the deterministic dynamics of the motif rate equations display a signature of the spontaneous symmetry breaking in sequence space, which occurs in the stochastic strand dynamics in a certain parameter regime. Another interesting future direction is to establish a systematic parameter inference from experimental data using the motif rate equations.

ACKNOWLEDGEMENTS

This project was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC-2094-390783311, ORIGINS) and via the CRC/TRR 392 Molecular Evolution (Project ID 521256690).

AUTHOR CONTRIBUTIONS

J.H.-K. performed the research. J.H.-K., T.A.E. and U.G. designed the research, with input from L.B. T.G. provided the code and help with the strand reactor simulations. J.H.-K., T.A.E. and U.G. wrote the paper, with input from all authors.

Appendix A: Motif Rate Equation Terms

In the following, we discuss the individual terms of the motif rate equations in detail. Let $\mathcal{M}_\ell$ be the space of all sequence motifs with a length of ℓ nucleotides. Mathematically, the motif rate equation terms depend on the motif concentrations that we store in the motif concentration vector $\vec{c}$, element of the *motif concentration space* $\mathbb{R}_{\geq 0}^{|\mathcal{M}_\ell|}$ with semi-positive real values and dimension of the number of motifs equal to the cardinality of the motif space $|\mathcal{M}_\ell|$. Every motif rate equation term indicated by an index i maps between identical motif concentration spaces: $\vec{f}_i : \mathbb{R}_{\geq 0}^{|\mathcal{M}_\ell|} \to \mathbb{R}_{\geq 0}^{|\mathcal{M}_\ell|}, \vec{c} \mapsto \vec{f}_i(\vec{c})$.

1. Motif Production Rate Calculation

The motif production rate becomes zero if the reaction is not allowed by our convention, i. e.if the second nucleotide of one reactant or the product is empty (zero), or in case of an empty spot between two nucleotides. If our convention is fulfilled, the motif production rate of a specific reaction is given by the product of the corresponding motif production rate constant multiplied by the concentrations of the reactants and template, as in Equation (6). This implies that we need to introduce an exceptional equation in case the left strand is a monomer, $l = (0, l_2, 0, 0)$. In this case, there is not left product, as in Equation (3), and the central product becomes

$$p_\mathrm{c} := p_\mathrm{c}(l, r) := (0, l_2, r_2, r_3). \tag{A1}$$

In case the right reacting strand is a monomer, Equation (4) still holds, but there is no right produced motif as in Equation (5).

The boundaries for the sums in the motif production terms for produced motifs follow from our convention. Accordingly, the first and last nucleotides of each motif can be empty or a letter, $l_1, r_4, t_1, t_4 \in \mathcal{A}_0$. The second nucleotide, as well as the third nucleotide of the template, is always a letter, $l_2, r_2, t_2, t_3 \in \mathcal{A}$. The third nucleotide of each motif is only empty if the motif is a monomer, i. e. the first and the fourth nucleotide are empty, $l_3 \in \mathcal{A} \cup (\{l_1\} \cap \{l_4\})$, $r_3 \in \mathcal{A} \cup (\{r_1\} \cap \{r_4\})$.

The motif production rate constants are approximated by the extension rate constants discussed next. To ensure positivity, we use logconcentrations to compute the motif production terms, the details of their computation are discussed in Appendix A 3.

2. Extension Rate Constants

We approximate the motif production rate constants by extension rate constants k_{ext} assuming that the motifs fully capture the participating strands. Thus, a produced four-nucleotide motif is assumed to be a tetramer, a beginning or ending motif a trimer, etc.. As in [7], the extension rate constants are given by the ligation rate constant (k_{lig}) devided by the dissociation constant of the complex C resulting from the hybridization of produced and templating motif,

$$k_{\text{ext}}(C) = \frac{k_{\text{lig}}(C)}{K_{\text{D}}(C)}. \tag{A2}$$

The ligation rate constant is given by a sequence independent ligation rate constant k_{lig_0} for perfect matches multiplied by the stalling factors $\Phi_{\pm}$.

$$k_{\text{lig}} = k_{\text{lig}_0}\Phi_+\Phi_-. \tag{A3}$$

The stalling factors are the product of the stalling parameters for mismatches close to the ligation spot.

$$\Phi_+ = \begin{cases} 1 & \text{for only perfect matches right of the ligation spot} \\ \sigma_1 & \text{for a mismatch directly right at the ligation spot} \\ \sigma_2 & \text{for a mismatch at the right next-nearest neighbour} \\ \sigma_1\sigma_2 & \text{for a mismatch at both nearest neighbours right of the ligation spot.} \end{cases} \tag{A4}$$

Φ_- analogously for the mismatches left of the ligation spot.

The dissociation constant is calculated from the hybridization energy difference, as defined in Equation (1). The hybridization energy is given by the sum of hybridization energies of the two by two nucleotide segments in the complex, where dangling ends have a hybridization energy of $\epsilon_{\text{com}} \pm \delta_\epsilon/2$ if they are perfectly matching and their sequence is alternating (+) or homogeneous (−), and ϵ_{1nc} if there is one mismatch, fully hybridized segments have a hybridization energy of $\overline{\gamma}_{\text{com}} \pm \Delta\gamma/2$ if they are perfectly matching and their sequence is alternating (+) or homogeneous (−), and γ_{1nc} (γ_{2nc}) if they have one (two) mismatches. The values of the reaction parameters are taken from [7] and stated in Table II.

Parameter	k_{lig_0}	ϵ_{com}	ϵ_{1nc}	$\overline{\gamma}_{\mathrm{com}}$	γ_{1nc}	γ_{2nc}
Value	$3.73 \cdot 10^{-6}$	-0.625	0.375	-1.25	0.375	0.75

TABLE II. Reaction parameters of the strand reactor simulation.

3. Motif Rate Equations with Log-Concentrations

To ensure that the concentrations stay non-negative, we express the differential equations in terms of the *log-concentrations* $\mu_m := \ln\left(\frac{c_m}{c_0}\right)$ and choose very low concentrations for non-occupied motifs. The corresponding vector is $\vec{\mu}$ with log-concentrations of each corresponding motif as entries $[\vec{\mu}]_m := \mu_m$. The motif rate equation is given by

$$\dot{\vec{\mu}} =: \tilde{\vec{f}}_{\mathrm{lig}}(\vec{\mu}) + \tilde{\vec{f}}_{\mathrm{break}}(\vec{\mu}). \tag{A5}$$

The individual terms transform to

$$\left[\tilde{\vec{f}}(\vec{\mu})\right]_m = \dot{\mu}_m = e^{-\mu_m}\partial_\tau \underbrace{e^{\mu_m}}_{c_m} = e^{-\mu_m}\left[\vec{f}(\vec{c}(\vec{\mu}))\right]_m, \qquad \tilde{\vec{f}} \in \{\tilde{\vec{f}}_{\mathrm{lig}}, \tilde{\vec{f}}_{\mathrm{break}}\}, \tag{A6}$$

where $[\vec{c}(\vec{\mu})]_m := e^{\mu_m}$.

The total motif production rate in terms of log-concentrations then becomes

$$\Lambda_{l,r,t} = \tilde{k}_{\mathrm{ext}\,p_{\mathrm{c}}(l,r),t} e^{\mu_l + \mu_r + \mu_t}. \tag{A7}$$

As before, we split the motif production terms into terms for the left reactant $\tilde{\vec{f}}_{\mathrm{lig_l}}$, the right reactant $\tilde{\vec{f}}_{\mathrm{lig_r}}$, and the three resulting motifs $\tilde{\vec{f}}_{\mathrm{lig_p}}$. Both, the ending as well as the leaving strand, are consumed in this reaction. This can be expressed in terms of log concentrations.

$$\left[\tilde{\vec{f}}_{\mathrm{lig_l}}(\vec{\mu})\right]_l = -e^{-\mu_l}\sum_{r,t}\Lambda_{l,r,t}, \tag{A8}$$

$$\left[\tilde{\vec{f}}_{\mathrm{lig_r}}(\vec{\mu})\right]_r = -e^{-\mu_r}\sum_{l,t}\Lambda_{l,r,t}. \tag{A9}$$

For the three produced motifs $(p_{\mathrm{l}}, p_{\mathrm{c}}, p_{\mathrm{r}})$, we get

$$\left[\tilde{\vec{f}}_{\mathrm{lig_p}}\right]_{p''} = \sum_{r_3,l_4,t}\Lambda_{l,r,t}\delta_{p'',p_{\mathrm{l}}(l,r)}e^{-\mu_{p_{\mathrm{l}}}} + \sum_{l_1,r_4,t}\Lambda_{l,r,t}\delta_{p'',p_{\mathrm{c}}(l,r)}e^{-\mu_{p_{\mathrm{c}}}} + \sum_{l_1,l_2,t}\Lambda_{l,r,t}\delta_{p'',p_{\mathrm{r}}(l,r)}e^{-\mu_{p_{\mathrm{r}}}}, \tag{A10}$$

with the exceptional case of monomers, which do not produce left (or right) products according to our convention, see Appendix A 1.

The breakage terms for log-concentrations split into terms for breaking ($\tilde{\vec{f}}^{-}_{\text{break}}$) and resulting ($\tilde{\vec{f}}^{+}_{\text{break}}$) motifs,

$$\tilde{\vec{f}}_{\text{break}}(\vec{\mu}) = \tilde{\vec{f}}^{-}_{\text{break}}(\vec{\mu}) + \tilde{\vec{f}}^{+}_{\text{break}}(\vec{\mu}). \tag{A11}$$

The terms for breaking motifs are

$$\left[\tilde{\vec{f}}^{-}_{\text{break}}(\vec{\mu})\right]_m = -e^{-\mu_m} \sum_{i=1}^{3} \tilde{k}_{\text{cut}}{}^{(i)}_{m} e^{\mu_m} \tag{A12}$$

$$= -\sum_{i=1}^{3} \tilde{k}_{\text{cut}}{}^{(i)}_{m}. \tag{A13}$$

The terms of the resulting motifs

$$\left[\tilde{\vec{f}}^{+}_{\text{break}}(\vec{\mu})\right]_{m=(m_1,m_2,m_3,0)} = +\sum_{x\in\mathcal{A}} \tilde{k}_{\text{cut}}{}^{(3)}_{(m_1,m_2,m_3,x)} e^{\mu_{(m_1,m_2,m_3,x)}-\mu_{(m_1,m_2,m_3,0)}} \tag{A14}$$

$$\left[\tilde{\vec{f}}^{+}_{\text{break}}(\vec{\mu})\right]_{m=(0,m_2,m_3,m_4)} = +\sum_{x\in\mathcal{A}} \tilde{k}_{\text{cut}}{}^{(1)}_{(x,m_2,m_3,m_4)} e^{\mu_{(x,m_2,m_3,m_4)}-\mu_{(0,m_2,m_3,m_4)}}, \tag{A15}$$

with all other entries being zero.

Summing all contributions together gives the overall motif rate equations for log-concentrations.

Appendix B: Monomer Dimer System

In the following, we test the motif rate equation by comparing the integrated solution to the analytic solution of a simple monomer dimer system with one letter in the alphabet.

Let m_1 denote a monomer sequence, m_2 a dimer. The total mass of the system, i. e.the total number of nucleotides M, is then

$$M = c_{m_1} + 2c_{m_2}. \tag{B1}$$

Thus, we can calculate the dimer concentration from the monomer concentrations and the motif rate equation reduces to one equation.

$$\dot{c}_{m_1} = -2\tilde{k}_{\text{ext}m_2,m_2} c_{m_1}^2 c_{m_2} \tag{B2}$$

$$= -\tilde{k}_{\text{ext}m_2,m_2} c_{m_1}^2 (c_{m_1} - M). \tag{B3}$$

We introduce the *monomer frequency* $\hat{c}_m := \frac{c_m}{M}$. Its time evolution is given by

$$\dot{\hat{c}}_{m_1} = -\tilde{k}_{\mathrm{ext}m_2,m_2} M^2 \hat{c}_{m_1}^2 (\hat{c}_{m_1} - 1). \tag{B4}$$

Integrating gives

$$\begin{aligned}
\tau_1 - \tau_0 &= \int_{\tau_0}^{\tau_1} \mathrm{d}\tau && \text{(B5)}\\
&= \int_{\hat{c}_{m_1}(t_0)}^{\hat{c}_{m_1}(\tau_1)} \frac{1}{\tilde{k}_{\mathrm{ext}m_2,m_2} M^2 \hat{c}_{m_1}^2 (1 - \hat{c}_{m_1})} \mathrm{d}\hat{c}_{m_1} && \text{(B6)}\\
&= \frac{1}{\tilde{k}_{\mathrm{ext}m_2,m_2} M^2} \left[\frac{1}{\hat{c}_{m_1}} + \ln(1 - \hat{c}_{m_1}) - \ln(\hat{c}_{m_1}) \right] + \text{IntegrationConstant}. && \text{(B7)}
\end{aligned}$$

The IntegrationConstant is given by the initial conditions at $\tau_1 = \tau_0$: $\hat{c}_{m_1}(\tau_0)$,

$$\text{IntegrationConstant} = -\frac{1}{\tilde{k}_{\mathrm{ext}m_2,m_2} M^2} \left[\frac{1}{\hat{c}_{m_1}(t_0)} + \ln(1 - \hat{c}_{m_1}(\tau_0)) - \ln(\hat{c}_{m_1}(\tau_0)) \right]. \tag{B8}$$

This way, we get the full solution

$$\tau_1 - \tau_0 = \frac{1}{\tilde{k}_{\mathrm{ext}m_2,m_2} M^2} \left[\frac{1}{\hat{c}_{m_1}} - \frac{1}{\hat{c}_{m_1}(\tau_0)} + \ln(1 - \hat{c}_{m_1}) - \ln(1 - \hat{c}_{m_1}(\tau_0)) - \ln(\hat{c}_{m_1}) + \ln(\hat{c}_{m_1}(\tau_0)) \right]. \tag{B9}$$

This case is equivalent to solving the motif rate equationswith motif production rate constants $\tilde{k}_{\mathrm{ext}p_c,t} = \delta_{p_c,m_2} \delta_{t,m_2} \tilde{k}_{\mathrm{ext}m_2,m_2}$. The resulting motif concentration trajectories coincide with the numerical solution to the motif rate equations, see Figure 12, validating our implementation.

Appendix C: Observables

1. Mass Conservation

A good test for the quality of the motif rate equations discussed in Section is mass conservation. In a scenario without influx nor outflux, nucleotides do not get lots, they just get rearranged within motifs. The test is nontheless nontrivial, as a single nucleotide can be part of several motifs, and only the concentration of the latter are followed by the simulation.

Let us call the *total mass* M of the system the concentration of nucleotides in the system and the mass $M(a)$ of a specific nucleotide $a \in \mathcal{A}$ the total concentration of nucleotides a.

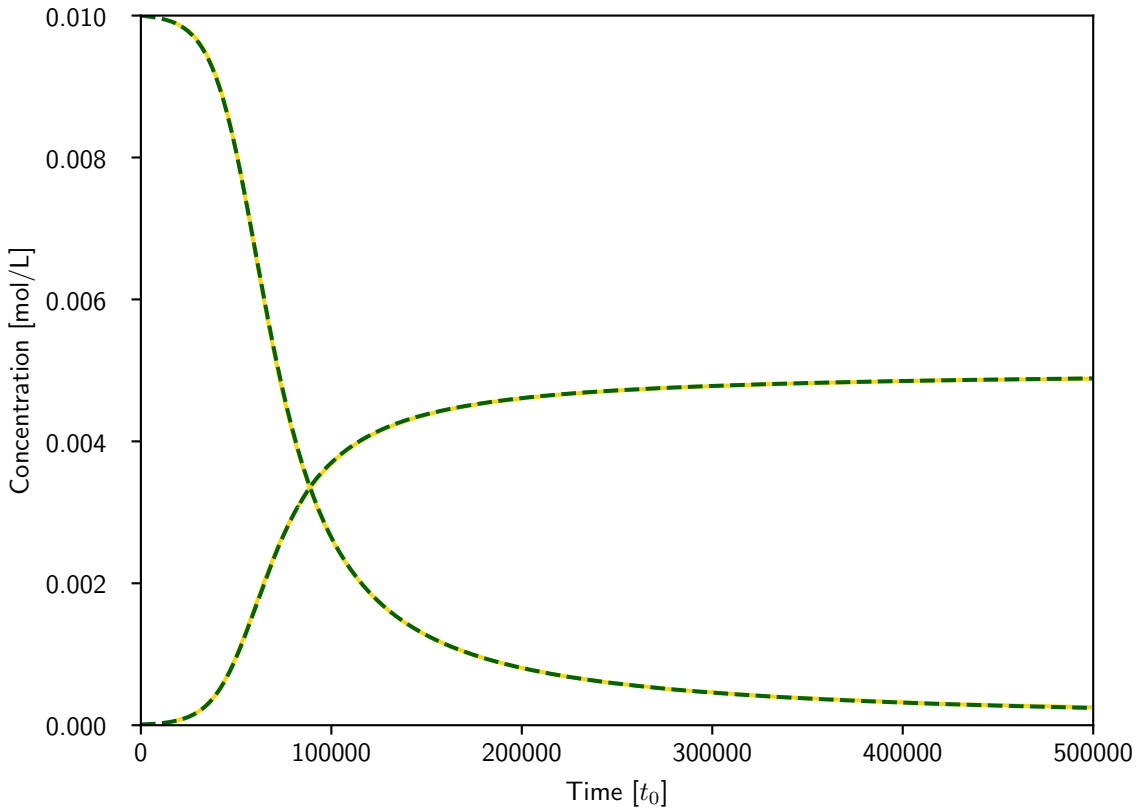

FIG. 12. Monomer (decreasing) and dimer (increasing) concentration trajectories determined analytically (solid gold) and numerically (dashed green).

The total mass can be computed by the sum over the mass of all nucleotides. We compute the nucleotide specific mass by scanning all motifs for the specific letter and adding the mass accordingly. There, we have to carefully consider our convention to prevent double counting.

According to our convention, the second nucleotide spot of every motif is occupied. It always fully contributes to the specific mass. The first and the last nucleotides never contribute to the mass, since they are tracked in the motifs before or after the current motif. The third letter only contributes, if the last nucleotide is empty, since the motif that would follow does not fulfill our convention and therefore is not tracked.

Thus, we have the following expression for the different nucleotide masses:

$$M(a) = \sum_{b_1,b_3,b_4 \in \mathcal{A}_0} c\left((b_1, a, b_3, b_4)\right) + \sum_{b_1 \in \mathcal{A}_0, b_2 \in \mathcal{A}} c\left((b_1, b_2, a, 0)\right), a \in \mathcal{A}. \tag{C1}$$

Those are separately conserved for zero influx and outflux.

2. Onset of growth in a Finite System

Here, we will estimate the difference of the onset of growth for the motif rate equations and the strand reactor simulation due to different system sizes.

Göppel et al. [7] derived an estimate for the onset of growth

$$\tau_{\text{est}} = \frac{\log\left[M^2 k_{\text{ext}}[1,1,2] - k_{\text{cut}}\right] - \log\left[M c_2(\tau=0) k_{\text{ext}}[1,2,2]\right]}{M^2 k_{\text{ext}}[1,1,2] - k_{\text{cut}}}. \tag{C2}$$

given the total number of nucleotides M and the sequence averaged extension rate constants,

$$k_{\text{ext}}[1,1,2] = \sum_{\substack{m_1,m_2\in\mathcal{A},\\ m_3\in\mathcal{A}^2}} k_{\text{ext}}[m_1,m_2,m_3], \tag{C3}$$

and approximating that the dimer concentration c_2 grows exponentially in the initial growth phase,

$$c_2(\tau) \approx c_2(\tau=0)\exp\left[\left([c_1(0)]^2 k_{\text{ext}}[1,1,2] - k_{\text{cut}}\right)\tau\right]. \tag{C4}$$

In a finite system, the onset of growth might be delayed, if the first dimer is produced after τ_{est}. Let c_Δ be the concentration of a single strand and τ_{dim}, the time at which the first dimer is produced. We use the exponential growth approximation for the initial dimer concentrations,

$$c_\Delta + c_2(\tau=0) = c_2(\tau_{\text{dim}}) \tag{C5}$$

$$\approx c_2(\tau=0)\exp\left[\left([c_1(0)]^2 k_{\text{ext}}[1,1,2] - k_{\text{cut}}\right)\tau_{\text{dim}}\right], \tag{C6}$$

and solve for τ_{dim},

$$\tau_{\text{dim}} = \frac{\log\left[1 + \frac{c_\Delta}{c_2(\tau=0)}\right]}{[c_1(0)]^2 k_{\text{ext}}[1,1,2] - k_{\text{cut}}}. \tag{C7}$$

If the initial pool consists mainly of monomers, the onset of growth gets delayed, if

$$\log\left[M^2 k_{\text{ext}}[1,1,2] - k_{\text{cut}}\right] - \log\left[M c_2(\tau=0) k_{\text{ext}}[1,2,2]\right] < \log\left[1 + \frac{c_\Delta}{c_2(t=0)}\right]. \tag{C8}$$

Solving for c_Δ gives

$$c_\Delta > c_2(\tau=0)\left(\frac{M^2 k_{\text{ext}}[1,1,2] - k_{\text{cut}}}{M c_2(\tau=0) k_{\text{ext}}[1,2,2]} - 1\right). \tag{C9}$$

For parameter set 0, the onset of growth is at $3.3\cdot 10^9 \tau_\Delta$, while the first dimer is produced before $2\cdot 10^7 \tau_\Delta$. Thus, the onset of growth is not delayed due to the finite size.

3. Mean Strand Length

While we can explicitly determine the mean strand length in the strand reactor simulation by averaging over the strand lengths from the output of the simulation, we need an estimate of it for the motif rate equations. For this, the length of all explicitly captured strands – the monomers and dimers – contribute directly. The mean length of longer strands needs to be estimated from the concentration of beginnings, continuations and ends. Beginnings contribute by one nucleotide to the mean length as their third and fourth nucleotide is considered by corresponding continuations or ends. For the same reasons, also continuations contribute to the mean length with one nucleotide. Ends contribute to the mean length with two nucleotides, since the second and the third nucleotide are not considered by other motifs. The total number of strands, $\mathcal{N}_{\text{strands}}$, is equal to the number of explicitly captured strands and ends. [29]

$$\mathcal{N}_{\text{strands}} = \frac{\sum_{m_1,m_3\in\mathcal{A}_0,m_2\in\mathcal{A}} c_{(m_1,m_2,m_3,0)}}{c_\Delta}. \tag{C10}$$

By taking the ends and considering that explicitly captured strands also have an empty spot at the end, the *mean strand length* $\overline{L}$ reduces to

$$\begin{aligned}\overline{L} &= \left(\sum_m c_m + \sum_{m_1\in\mathcal{A}_0,m_2,m_3\in\mathcal{A}} c_{(m_1,m_2,m_3,0)}\right)\Big/\mathcal{N}_{\text{strands}} \\ &= 1 + \frac{\sum_m c_m - \sum_{m_1} c_{(0,m_1,0,0)}}{\mathcal{N}_{\text{strands}}}.\end{aligned} \tag{C11}$$

[1] Adriana Calacca Serrão, Felix T Dänekamp, Zsófia Meggyesi, and Dieter Braun. Replication elongates short DNA, reduces sequence bias and develops trimer structure. *Nucleic Acids Research*, 52(3):1290–1297, 12 2023.

[2] F.H.C. Crick. The origin of the genetic code. *Journal of Molecular Biology*, 38(3):367–379, 1968.

[3] 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*, 40(10):4711–4722, 02 2012.

[4] Harold Fellermann, Shinpei Tanaka, and Steen Rasmussen. Sequence selection by dynamical symmetry breaking in an autocatalytic binary polymer model. *Phys. Rev. E*, 96:062407, Dec 2017.

[5] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.

[6] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[7] Tobias Göppel, Joachim H. Rosenberger, Bernhard Altaner, and Ulrich Gerland. Thermodynamic and kinetic sequence selection in enzyme-free polymer self-assembly inside a non-equilibrium rna reactor. *Life*, 12(4), 2022.

[8] Paul G Higgs and Niles Lehman. The RNA world: molecular cooperation at the origins of life. *Nat Rev Genet*, 16(1):7–17, November 2014.

[9] Gabin Laurent, Tobias Göppel, David Lacoste, and Ulrich Gerland. Emergence of homochirality via template-directed ligation in an rna reactor. *PRX Life*, 2:013015, Mar 2024.

[10] Kevin Leu, Eric Kervio, Benedikt Obermayer, Rebecca M. Turk-MacLeod, Caterina Yuan, Jesus-Mario Jr. Luevano, Eric Chen, Ulrich Gerland, Clemens Richert, and Irene A. Chen. Cascade of reduced speed and accuracy after errors in enzyme-free copying of nucleic acid sequences. *Journal of the American Chemical Society*, 135(1):354–366, 01 2013.

[11] Gabrielle Leveau, Daniel Pfeffer, Bernhard Altaner, Eric Kervio, Franziska Welsch, Ulrich Gerland, and Clemens Richert. Enzyme-free copying of 12 bases of rna with dinucleotides. *Angewandte Chemie International Edition*, 61(29):e202203067, 2022.

[12] Yoshiya J. Matsubara and Kunihiko Kaneko. Kinetic selection of template polymer with complex sequences. *Phys. Rev. Lett.*, 121:118101, Sep 2018.

[13] Since this depends on the concentration of motifs, it is not a rate constant as such but it still is not the rate of the breakage as the concentration of the reactants is still missing. We thus call it effective rate constant.

[14] We could also consider the number of beginnings instead, but for consistency, the number of ends is equal to the number of beginnings as every strand that begins also ends.

[15] Benedikt Obermayer, Hubert Krammer, Dieter Braun, and Ulrich Gerland. Emergence of information transmission in a prebiotic rna reactor. *Phys. Rev. Lett.*, 107:018101, Jun 2011.

[16] L.E. Orgel. Evolution of the genetic apparatus. *Journal of Molecular Biology*, 38(3):381–393, 1968.

[17] Abe Pressman, Celia Blanco, and Irene A. Chen. The rna world as a model system to study the origin of life. *Current Biology*, 25(19):R953–R963, 2015.

[18] Sudha Rajamani, Justin K. Ichida, Tibor Antal, Douglas A. Treco, Kevin Leu, Martin A. Nowak, Jack W. Szostak, and Irene A. Chen. Effect of stalling after mismatches on the error catastrophe in nonenzymatic nucleic acid replication. *Journal of the American Chemical Society*, 132(16):5880–5885, 04 2010.

[19] Joachim H. Rosenberger, Tobias Göppel, Patrick W. Kudella, Dieter Braun, Ulrich Gerland, and Bernhard Altaner. Self-assembly of informational polymers by templated ligation. *Phys. Rev. X*, 11:031055, Sep 2021.

[20] Jack W. Szostak. The eightfold path to non-enzymatic rna replication. *Journal of Systems Chemistry*, 3(1):2, 2012.

[21] Alexei V. Tkachenko and Sergei Maslov. Spontaneous emergence of autocatalytic information-coding polymers. *The Journal of Chemical Physics*, 143(4):045102, 2015.

[22] Alexei V. Tkachenko and Sergei Maslov. Onset of natural selection in populations of autocatalytic heteropolymers. *The Journal of Chemical Physics*, 149(13):134901, 2018.

[23] Shoichi Toyabe and Dieter Braun. Cooperative ligation breaks sequence symmetry and stabilizes early molecular replication. *Phys. Rev. X*, 9:011056, Mar 2019.

[24] Andrew S. Tupper and Paul G. Higgs. Rolling-circle and strand-displacement mechanisms for non-enzymatic rna replication at the time of the origin of life. *Journal of Theoretical Biology*, 527:110822, 2021.

[25] Andrew S. Tupper, Kevin Shi, and Paul G. Higgs. The role of templating in the emergence of rna from the prebiotic chemical mixture. *Life*, 7(4), 2017.

[26] Nilesh Vaidya, Sara Imari Walker, and Niles Lehman. Recycling of informational units leads to selection of replicators in a prebiotic soup. *Chemistry & Biology*, 20(2):241–252, 2013.

[27] Petr Šulc, Flavio Romano, Thomas E. Ouldridge, Jonathan P. K. Doye, and Ard A. Louis. A nucleotide-level coarse-grained model of RNA. *The Journal of Chemical Physics*, 140(23):235102, 06 2014.

[28] Since this depends on the concentration of motifs, it is not a rate constant as such but it still is not the rate of the breakage as the concentration of the reactants is still missing. We thus

call it effective rate constant.

[29] We could also consider the number of beginnings instead, but for consistency, the number of ends is equal to the number of beginnings as every strand that begins also ends.