Onset of autocatalysis of information-coding polymers

Alexei V. Tkachenko\textsuperscript{1} and Sergei Maslov\textsuperscript{2}

\textsuperscript{1}Center for Functional Nanomaterials, Brookhaven National Laboratory, Upton NY 11973

\textsuperscript{2}Biological, Environmental and Climate Sciences Department, Brookhaven National Laboratory, Upton NY 11973

November 11, 2014

Self-replicating systems based on information-coding polymers are of crucial importance in biology. They also recently emerged as a paradigm in design on nano- and micro-scales. We present a general theoretical and numerical analysis of the problem of spontaneous emergence of autocatalysis for heteropolymers capable of template-assisted ligation driven by cyclic changes in the environment. Our central result is the existence of the first order transition between the regime dominated by free monomers and that with a self-sustaining population of sufficiently long oligomers. We provide a simple mathematically tractable model that predicts the parameters for the onset of autocatalysis and the distribution of chain lengths, in terms of monomer concentration, and two fundamental rate constants. Another key result is the emergence of the kinetically-limited optimal overlap length between a template and its two substrates. Template-assisted ligation allows for heritable transmission of information encoded in oligomer sequences thus opening up the possibility of long-term memory and evolvability of such systems.

Life as we know it today depends on replication of information-coding polymers. Its emer-
gence from non-living matter is one of the greatest mysteries of fundamental science. In addition, self-replicating systems recently emerged as an exciting new direction in studies of nano- and meso-scale physics potentially leading to engineering applications. The central challenge in both of these fields is to come up with a simple physically-realizable system obeying laws of thermodynamics, yet ultimately capable of Darwinian evolution when it is subject to non-equilibrium driving forces. Chemical networks of generic molecules that are not individually self-replicating but as a group engaged in mutual catalysis is a popular candidate for such a system. Heteropolymers with the property of sequence-specific reversible binding (hybridization) such as nucleic acids provide an attractive example of such systems. Their advantages are twofold. First, by their nature they are information-coding, and, second, their hybridization provides a specific chemically plausible mechanism for mutual catalysis. In fact, the most successful experimental realization of an autonomous self-replicating system involves a set of mutually catalyzing RNA-based enzymes (ribozymes) that show evolution-like behavior. This is viewed as a major evidence for RNA-world hypothesis (see e.g. Refs. [9–11]). However, the ribozyme activity requires relatively long (hundreds of nucleotides) and carefully designed sequences.

Long polymers can be generated e.g. by traditional reversible step-growth polymerization that combines random concatenation and fragmentation of polymer chains. It has been recently theoretically and experimentally demonstrated that in this type of process the polymer length can be drastically increased in non-equilibrium situations such as temperature gradient. However, even when long chains are formed the probability of spontaneous emergence of a sequence with enzymatic activity remains vanishingly small due to the exponentially large number of all possible
sequences.

Thus there is a strong need for a mechanism that combines the emergence of long chains with dramatic reduction of informational entropy of the sequence population. A promising candidate for such mechanism is provided by template-assisted ligation. In this process pairs of polymers are brought together by hybridization with a complementary polymer chain serving as the template and eventually ligated to form a longer chain. Unlike non-templated reversible step-growth polymerization used in Ref.\[12\], this mechanism naturally involves the transmission of information from a template to the newly ligated chain, thus opening an exciting possibility of long-term memory and evolvability. An early model involving template-assisted polymerization was proposed by P. W. Anderson and colleagues\[13,14\]. It also has been subject of more recent experimental and theoretical studies\[15,17\]. In particular, in Ref.\[15\] it has been demonstrated that, for a specific choice of parameters, a combination of non-template and template-assisted ligation can lead to the emergence of long (100 monomer) oligonucleotides.

In this work we carried out theoretical and numerical analysis of a generic system in which polymerization is driven solely by template-assisted ligation. Unlike models with significant contribution of non-templated concatenation, the emergence of long chains in our system represents a non-trivial chicken-or-egg problem. Indeed, the formation of long chains depends on the presence of other chains serving as templates. In the regime where long chains do emerge, template-assisted polymerization provides built-in mechanism for the transmission of inheritable information. While in this work we study the problem in the approximation of random sequence composition it lays
the foundation for future work exploring the evolution in the sequence space.

**Results.**

**Description of the model.** Our model explores the fundamental problem of the emergence of a self-sustaining population of information-coding polymers out of the “primordial soup” of their monomers. It is driven out of equilibrium by cyclic changes in physical conditions such as temperature, salt concentration, pH, etc. (see Fig. [1]ab). Polymerization occurs during the ”night” phase of each cycle when the existing heteropolymers serve as templates for formation of progressively longer chains by catalyzing ligation of shorter polymers. During the ”day” phase of each cycle all multi-chain structures separate and the system returns to the state of dispersed individual polymers.

Motivated by the example of nucleic acids, we consider a general case of information-coding heteropolymers composed of $z$ types of monomers capable of making $z/2$ mutually complementary pairs. For example, RNA is made of $z = 4$ monomers forming 2 complementary pairs $A - U$ and $C - G$ responsible for double-stranded RNA structure. Similarly, we assume that hybridization between complementary segments of our generalized polymers results in formation of a double-stranded structure. The key process in our model happening only during the night phase of each cycle is the formation of hybridized complexes joining together two or more chains. The ligation takes place in special type of three-polymer hybridized complexes shown in the right side of Fig. [1]b. These complexes position the end groups of two polymers referred to as ”substrates” $S_1$ and $S_2$ next to each other by the virtue of their hybridization with the third chain referred to as
"template" $T$. Once the substrates are properly positioned the new bond joining them together is formed at a constant rate. Generally speaking the ligation of two chains could happen even without template$^{12,15}$. However, throughout this work we assume that the rate of such spontaneous mergers is negligible compared to the rate of template-assisted ligation - the limit the least favorable for polymerization. We further assume that each of the intra-polymer bonds can spontaneously break at a constant rate making the overall fragmentation rate of a chain proportional to its length.

If one was to leave a mixture of polymers in the night phase long enough, hybridization of multiple chains will result in formation of a gel-like aggregate shown in Fig. 1c, effectively stopping ligation. The "day" phase of the cycle (Fig. 1a) when all structures of hybridized polymers dissociate while keeping their stronger internal bonds intact plays the role of the "reset" returning the system to a mixture of free polymers ready for the next night phase of the cycle.

**Two timescales of the model.** Dynamical processes in our model occur on two distinct timescales. On a shorter scale of a single night, polymers hybridize with each other forming various multi-chain complexes (including the three-body ones shown in Fig. 1a catalyzing subsequent ligation of polymer chains). The likelihood of hybridization strongly depends on the overlap length $k$ - the number of consecutive mutually complementary monomers in a pair of interacting chains. From thermodynamics point of view, longer overlaps result in more stable complexes thus increasing the likelihood of ligation. However, the process is kinetically limited: the longer is $k$ the smaller is the probability of finding a complementary segment, which for random chain sequences is proportional to $z^{-k}$. This exponential factor dramatically reduces the effective hybridization rate for large $k$. 
Figure 1: The schematic representation of fundamental processes in our system. a) The "day" phase during which all hybridized complexes between heteropolymers dissociate and ligation completely stops, while fragmentation continues in all phases of the cycle. b) The "night" phase during which some polymer chains hybridize and then undergo template-assisted ligation. The ends of substrates $S_1$ (green) and $S_2$ (red) hybridized with a template $T$ (purple) are ligated at a constant rate with the newly formed bond shown in blue. c) If the "night" phase is sufficiently long heteropolymers enter the aggregation regime in which ligation effectively stops.
As a consequence of the competition between thermodynamic effects favoring large $k$ and kinetic effects opposing it, the optimal overlap length $k_0$ emerges (see Methods and SI for details). Thus hybridization processes in our system are dominated by $k \leq k_0$. The overlap $k_0$ increases during the night phase with its final value set by either the end of the night or (in case of long nights) by the onset of the aggregation regime at which most chains are immobilized in a gel-like structure (Fig. 1c).

In this study we focus primarily on the dynamics taking place over multiple day/night cycles. The main input parameter from the intra-night (short timescale) kinetics to the multi-cycle (long timescale) dynamics is the hybridization cutoff length $k_0$. The slow dynamics can be described in terms of time-averaged ligation and fragmentation rates, $\lambda$ and $\beta$, respectively. We define $\lambda$ as the rate of bond formation provided that the ends of two substrates are properly positioned next to each other due to hybridization with the template. We further assume that the characteristic fragmentation time $1/\beta$ is much longer than the duration of the day-night cycle ensuring the separation between short and long timescales in the problem. Both $\lambda$ and $\beta$ are averaged over the duration of the day-night cycle with the understanding that fragmentation happens continuously throughout the cycle (possibly with different day- and night-rates), while ligation only occurs during the night phase. Thus $\lambda$ implicitly depends on relative durations of night and day phases.

Let $C$ be the overall monomer concentration including both free monomers and those bound in all chains. One of the major assumptions used in our study is the Random Sequence Approximation (RSA) according to which each monomer in every chain can be of any type with
equal probability $1/z$. On the one hand, the RSA greatly simplifies the problem and allows us to get a concise analytical solution. On the other hand, in order to understand the transmission of sequence-encoded information and the long-term memory in our system this approximation need to be relaxed in future studies. In the case of random sequence composition, the population of heteropolymers is fully characterized by their length distribution $f_l$, defined in such a way that $C \cdot f_l$ is the concentration of all polymers of length $l$. By this definition $f_l$ is subject to the normalization condition $\sum_{l=1}^{\infty} lf_l = 1$. Within the RSA the fraction of polymers with a specific sequence is given by $z^{-l} \cdot f_l$.

**Detailed balance approximation.** In what follows we study the system both numerically and analytically. For template-assisted ligation the effective two-polymer merger rate $\mu$ is given by the ligation rate $\lambda$ multiplied by the probability of hybridization of a template $T$ with two substrates $S_1$ and $S_2$ bring them into end-to-end configuration (see Fig. 1b for the illustration). The major step in constructing an approximate analytical solution of the problem is the assumption of a detailed balance between template-assisted ligation and fragmentation in the steady state of the system:

$$
\beta f_{l+m} = \mu f_l \cdot f_m.
$$

(1)

Here, the left-hand side describes the rate at which a chain of length $l + m$ breaks into two pieces of lengths $l$ and $m$ correspondingly. Conversely, the right-hand side is the effective merger rate (hybridization + ligation) at which polymers of lengths $l$ and $m$ are ligated to form a longer chain of length $l + m$. Note that according to this description the rate at which a polymer breaks into arbitrary two pieces is proportional to its length or rather its number of intra-polymer bonds.
The detailed balance approximation is not a priory justified in driven, non-equilibrium systems such as ours. However, for chains longer than the optimal overlap length $k_0$ defined above, the probability of hybridization with a template and thus the effective merger rate $\mu$ becomes independent of their lengths (see Methods for derivation and details). Once both $\mu$ and $\beta$ are independent of polymers’ lengths, our system becomes mathematically equivalent to the well known reversible step-like polymerization process for which the detailed balance approximation is satisfied by the virtue of laws of equilibrium thermodynamics.

In spite of this superficial similarity simplifying mathematical calculations, our system remains intrinsically non-equilibrium since the effective merger rate $\mu$ relies on hybridization between templates and substrates cycled through day and night phases as shown in Fig. [1]ab. In addition, the Eq. (1) is expected to break down for chains shorter than $k_0$. To validate our mathematical insights, the analytic solution shown below was followed by numerical simulations of the system without the detailed balance approximation. The agreement between our analytical and numerical results for polymers longer than $k_0$ further confirms the validity of our approach.

The Eq. (1) can be used to determine the steady state polymer length distribution $f_l$. Indeed, it is satisfied by the exponential distribution:

$$f_l = \frac{\beta}{\mu} \exp\left(-\frac{l}{\bar{L}}\right), \quad (2)$$

where the characteristic chain length, $\bar{L}$, is determined by the normalization condition $\sum_{l=1}^{\infty} l f_l = 1$ leading to $\frac{\beta}{\mu}\bar{L}^2 = 1$. This result was obtained by replacing the discrete sum above with the integral, which works in the limit $\bar{L} \gg 1$ (see SI for the exact formula in which this approximation
is relaxed). Hence, the characteristic chain length in the steady state exponential distribution is given by

\[ \bar{L} = \sqrt{\frac{\mu}{\beta}} \]

**(Autocatalytic regime.** As discussed above \( \mu \) is an effective two-polymer merger rate proportional to the probability of finding two terminal ends attached to a template followed by ligation. This probability in turn depends on (a) the overall concentration \( C \) and the length distribution of potential templates (b) the strength and kinetics of interactions between the complementary segments on a template and its two substrates. In general the interaction strength between any two chain segments increases with the overlap length \( k \) of the region over which they are complementary to each other. Here we assume a simple linear relationship in which the binding free energy is given by \( \Delta G_0 + k \cdot \Delta G \) where \( \Delta G \) is the (negative) binding free energy between two complementary monomers, while \( \Delta G_0 \) is the initiation free energy. It is straightforward to generalize this assumption to sequence-dependent interactions. For nucleic acids, when two chain segments are bound to the same template and are directly adjacent to each other there is an additional gain in free energy \( \Delta G_{st} \) due stacking.

In case of the RSA an increase in the interaction strength with \( k \) is offset by the decrease in the probability \( 1/z^k \) of finding perfectly complementary sequences of the template and its substrate. The net result is that the hybridization probability is proportional to \( \exp(k \cdot \epsilon) \), where \( \epsilon = -\Delta G/k_B T - \log(z) \) is the effective parameter combining thermodynamic and combinatorial factors. Template-assisted ligation happens at appreciable rates only for positive \( \epsilon \) realized
when $\Delta G < -k_B T \log(\varepsilon)$. Kinetic effects lead to a strong and abrupt drop in the hybridization probability for $k > k_0$ (see the section “Kinetics of hybridization” for details). By neglecting the contribution of overlap lengths longer than $k_0$ one gets

$$\mu = \lambda \left( \frac{C}{C_0} \right)^2 \sum_{k_1=1}^{k_0} \exp(k_1 \cdot \varepsilon) \sum_{k_2=1}^{k_0} \exp(k_2 \cdot \varepsilon) \sum_{l=k_1+k_2+1}^{\infty} (l - k_1 - k_2 + 1) \cdot f_l.$$  (4)

Here $\lambda$ is the previously introduced rate of formation of new bonds at conjugated ends, while $k_1$ and $k_2$ are the overlap lengths between the template and each of the two substrates, respectively. We also introduced the reference concentration $C_0 = \exp[-(\Delta G_0 + \Delta G_{st}/2)/k_B T]$ (in molar) accounting for initiation and stacking free energies. The term $(C/C_0)^2$ reflects the fact that template-assisted ligation is a three-body interaction involving two substrates and one template. The last sum in the r.h.s. of the Eq. (4) is equal to the probability of finding a template region of length $k_1 + k_2$ within a longer heteropolymer. It takes into account that a chain of length $l \geq k_1 + k_2$ has $l - k_1 - k_2 + 1$ sub-sequences of length $k_1 + k_2$. Requirements of sequence complementarity between the template and each of two substrates were absorbed into the definition of $\varepsilon$ within the RSA. Substituting the exponential distribution $f_l$ given by the Eq. (2), performing the triple summation in Eq. (4), and neglecting the terms $\sim 1/L$ but not $\sim k_0/L$ within the exponents approximately gives $\mu = \lambda(C/C_0)^2 \exp(2k_0 \cdot (\varepsilon - 1/L)) / [1 - \exp(-\varepsilon)]^2$ Substituting this expression into the Eq. (3) one gets the self-consistency equation for $L$:

$$L \exp \left( \frac{k_0}{L} \right) = \frac{C}{C_0} \cdot \sqrt{\frac{\lambda}{\beta}} \cdot \frac{\exp(k_0 \varepsilon)}{1 - \exp(-\varepsilon)},$$  (5)

(see the Eq. (S11) in the SI for a more cumbersome expression derived without the large $L$ approximation). The l.h.s. of this equation reaches its minimal value of $\varepsilon \cdot k_0$ at $L = k_0$. As a result,
the equation has solutions only for concentrations $C$ above a certain threshold value given by

$$C_{\text{down}} = k_0C_0 \sqrt{\frac{\beta}{\lambda}} \exp(1 - k_0 \epsilon) \cdot (1 - \exp(-\epsilon)) \quad (6)$$

For $C$ sufficiently above this threshold, one can neglect the exponential term in the l.h.s. of the Eq. \(S11\) so that the characteristic polymer length $\bar{L}$ linearly increases with the concentration as

$$\bar{L} = \frac{C}{C_0} \cdot \sqrt{\frac{\lambda}{\beta}} \cdot \frac{\exp(k_0 \epsilon)}{1 - \exp(-\epsilon)} \quad (7)$$

Conversely, for monomer concentrations $C$ below this threshold we don’t expect formation of long heteropolymers. This suggests the existence of a first-order transition between the regimes dominated by mostly free monomers and that with a self-sustaining population of long heteropolymeric chains.

**Onset of polymerization.** To verify and refine our predictions we approach this problem from the opposite limit where the system consists predominantly of monomers i.e. $f_1 \approx 1$. We now explore the stability of this monomer mixture with respect to formation of dimers. In this limit, dimer fraction $f_2$ obeys the following kinetic equation:

$$\frac{df_2}{dt} = -\beta f_2 + \lambda \left(\frac{C}{C_0}\right)^2 \exp(2\epsilon)f_1^2 f_2 \quad , \quad (8)$$

where the second term in the r.h.s. reflects the fact that a dimer can be formed out of two monomers and this process needs to be catalyzed by a complementary dimer. The critical concentration $C_{\text{up}}$ above which spontaneously formed dimers will exponentially self-amplify is given by

$$C_{\text{up}} = C_0 \sqrt{\frac{\beta}{\lambda}} \exp(-\epsilon) \quad . \quad (9)$$
Thus we confirm the existence of instability in a mixture of monomers at concentration $C > C_{\text{up}}$ with respect to template-assisted formation of longer chains. Note that, as expected for a first-order phase transition, the instability threshold $C_{\text{up}}$ (Eq. (9)) approached from below exceeds the instability threshold $C_{\text{down}}$ (Eq. (6)) approached from above. Thus the system will be hysteretic for $C_{\text{down}} < C < C_{\text{up}}$.

**Numerical results.** To check our calculations we carried out detailed numerical simulations of our system. Specifically we numerically solved a system of coupled kinetic equations describing template-assisted ligation and fragmentation processes and calculated the steady state distribution $f_l$ (see SI for details). *In our simulations we did not assume the detailed balance.*

The results of numerical simulations are in excellent agreement with our analytical calculations. For high enough concentrations $C$ the length distribution $f_l$ has a long exponential tail covering the region $l > k_0$. Chains of length shorter than $k_0$, which do not obey detailed balance, exhibit a much faster decay as a function of $l$ (see Fig. 2).

Our simulations also confirmed the existence of a first-order transition to a regime dominated by monomers as concentration $C$ was reduced (see the red line with circles in Fig. 3). The decay length $\bar{L}$ of the exponential tail of $f_l$ for $l \geq k_0$ plays the role of the order parameter in this transition. When plotted as a function of concentration $C$ in Fig. 3 it exhibits sharp discontinuities and hysteretic behavior as expected in a first-order phase transition. Our analytical results given by the Eq. (S11) (black dashed line in Fig. 3) are in a good agreement with our numerical simulations. The transitions from monomers to long-chained polymers and back in our numerical simulations
Figure 2: **Chain length distributions.** A set of chain length distributions $f_l$ plotted for different values of the control parameter $\Gamma = \frac{C}{c_0} \sqrt{\lambda/\beta}$ as found by numerical simulations with $k_0 = 3$ and $\epsilon = 1$. Distributions in the autocatalytic regime are characterized by long exponentially distributed tails for chains with $l > k_0$. Note a sharp transition between monomer-dominated and autocatalytic regimes.

Kinetics of hybridization As was discussed above, the optimal overlap length $k_0$ between a template and a substrate is determined by the competition between the opposing trends in hybridization
Figure 3: A hysteretic first order transition between monomer and autocatalytic regimes. Blue and red lines/symbols show the characteristic length $\bar{L}$ in our numerical simulations with $k_0 = 3$ for increasing (blue diamonds), and decreasing (red circles) concentration $C$, correspondingly. The dashed line is the prediction of our simplified model given by Eq. (S11). Arrows indicate $C_{\text{up}}$ and $C_{\text{down}}$ given by Eqs. (9) and (6) correspondingly.

and dissociation rates. On the one hand, within the RSA, the probability of finding a pair of polymers with complementary sequences of length $k$, and hence their hybridization rate exponentially decreases as $1/z^k$. On the other hand, the dissociation rate between a template and a substrate also
exponentially decreases with $k$ as $\exp(-k \cdot \Delta G/k_B T)$ due to greater thermodynamic stability of longer complementary duplexes. As was explained above for $\epsilon > 0$, the equilibrium determined by the ratio between hybridization and dissociation rates would favor the longest possible overlap. However, for a finite time window during which hybridization takes place, the probability to find a segment of length $k$ in a hybridized complex is strongly peaked as a function of $k$ (see Fig. 4 and SI for details). This peak slowly shifts with time towards larger $k$ and its position at the end of the night corresponds to the optimal overlap length $k_0$ used in calculations above.

Our model assumes cyclic changes between "day" and "night" phases. In the beginning of a night phase all polymers are unhybridized, but as time progresses they start forming duplexes of progressively longer lengths. The probability of finding any given segment in a duplex remains low at the early stage of this process. However, if the duration of the night phase is long enough, there will be a time point at which individual polymers will on average have around one hybridized partner. Note that a single polymer may simultaneously have more than one hybridized partner as long as duplexes with different partners do not overlap with each other. Around this time most polymers in our pool will become immobilized in a gel-like structure schematically depicted in Fig. 1c. In such aggregation phase $k_0$ saturates and the formation of new hybridized complexes effectively stops. Indirect experimental evidence for such aggregation was recently reported by Bellini et al.\textsuperscript{16}.

According to the analysis presented above, the characteristic chain length $\bar{L}$ given by Eq. (7) exponentially increases with $k_0$. In the presence of aggregation this growth is eventually arrested.
Figure 4: **Time evolution of the hybridization probability.** The probability that a segment of length \( k \) is hybridized to its complementary partner (Eq. (S6) in SI) is strongly peaked at \( k = k_0 \sim \log t \) (see Eq. (S7)). Different colors from red to violet correspond to linearly increasing times \( t \) since the beginning of the night phase of the cycle.

The upper bound on \( \bar{L} \) reached in this case can be determined self-consistently. Indeed, the aggregation starts when individual polymers on average have around one hybridized partner. A chain of length \( \bar{L} \gg k_0 \) contains \( \bar{L} - k_0 + 1 \simeq \bar{L} \) of segments of length \( k_0 \). The probability of each of these segments to be tied up in a duplex at any given time is \( \frac{C}{C_0} \exp(k_0 \cdot \epsilon) \). Thus the transition to the
aggregated state is expected when

\[
\frac{\bar{L} C}{C_0} \exp(k_0 \epsilon) \simeq 1.
\]  

(10)

Combining this expression with the Eq. (7) and ignoring factors of order of 1 one gets the upper bound \( \bar{L}_{\text{max}} \) on characteristic polymer length that can, in principle, be reached by increasing the duration of the night phase as

\[
\bar{L}_{\text{max}} \simeq \left( \frac{\lambda}{\beta} \right)^{\frac{1}{4}}.
\]  

(11)

**Discussion**

In conclusion, we considered a general case of random heteropolymers capable of template-assisted ligation. As such our model is applicable to both nucleic acids at the dawn of life as well as to artificial self-replicating nano- or micro- structures\(^I\). The major conclusions of our study are as follows. We demonstrated that a population of long chains can be sustained by mutual catalysis sustained exclusively by template-assisted ligation. This state is separated from the monomer-dominated one by a hysteretic first order phase transition (Eq. (6)) as a function of concentration. We also demonstrated that the template-assisted ligation in our system is dominated by contributions from template-substrate pairs complementary over a well-defined length \( k_0 \) that is kinetically limited. The average length of heteropolymers exponentially depends on \( k_0 \) with the upper bound given by a very simple expression, Eq. (11), depending only on the ratio between ligation and breakage rates.

The spontaneous emergence of long information-coding polymers demonstrated in our study
is of conceptual importance to the long-standing problem of the origin of life. Indeed, we offer a physically plausible path leading from the primordial soup dominated by monomers to a population of sufficiently long self-replicating chains where Darwinian evolutionary processes would in principle be able to kick in. This transition is one of the least understood processes in the RNA-world hypothesis. It is known that functional RNA-based enzymes (ribozymes) need to be sufficiently long which makes their spontaneous formation prohibitively unlikely. According to our analysis both the characteristic chain length and the minimal monomer concentration required for autocatalysis depend on the ratio of ligation and breakage rates. Large values of this ratio $\lambda/\beta \gg 1$ would allow long chains to form at physically possible concentrations $C \ll 1$M. One of the reasons that such spontaneous emergence of long-chained polymers has never been observed is that in experimental systems studied so far the ratio $\lambda/\beta$ remained low due to a very slow ligation process. Note that ligation and breakage processes in our system are not direct opposites of each other. Indeed, the ligation of e.g. nucleic acids requires activated terminal bases carrying free energy sufficient to form a new intra-polymer bond. To achieve the conditions necessary for our autocatalytic regime one needs to either use heteropolymers chemically different from modern nucleic acids or to develop new activation pathways different from what has been used in experiments so far. The ligation can be further assisted e.g. by absorption of polymers onto properly selected crystalline interfaces.

We demonstrated that the spontaneous emergence of long chains is possible even in the limit where direct (non-templated) bond formation is negligible. This is especially important since non-templated polymerization is a regular equilibrium phenomenon and as such has short memory.
In contrast, the transmission of sequence information by template-assisted ligation opens up an exciting possibility of long-term memory effects and ultimately of Darwinian evolution in the space of polymer sequences.

The present study was limited to the simplest version of the problem in which sequences of all heteropolymers were assumed to be completely random. Incorporation of sequence effects is the logical next step in the development of our model, which we are currently working on. There are several conceptually distinct yet non mutually exclusive scenarios giving rise to over-representation of certain sequences in the pool of heteropolymers. The first one is driven by sequence dependence of model parameters such as hybridization free energies, fragmentation and ligation rates, and monomer composition of the "primordial soup". The other scenario is the spontaneous symmetry breaking in the sequence space\(^{13,18}\). Specifically, our results obtained within the Random Sequence Approximation need to be checked for local and global stability. The local stability analysis deals with small deviations from a state in which populations of all sequences are equal, while global one perturbs the system by strongly over-representing a small subset of sequences. Such instability would signal a symmetry breaking and would provide a scenario for dramatic decrease in informational entropy of the population.

**Methods**

**\(k\)-mers and their hybridization dynamics** To describe hybridization dynamics during the night phase we introduce the concept of a \(k\)-mer defined as the segment of \(k\) monomers with the specific sequence \(\sigma\) within a longer chain of length \(l \geq k\). Let \(C \cdot p_k^{(\sigma)}\), be the concentration of \(k\)-mers
with particular sequence $\sigma$. Let $C \cdot P_k$ be the concentration of all $k$-mers of length $k$, regardless of their sequences. By definition, $P_k = \sum_\sigma p_k^{(\sigma)}$. If all the sequences are completely random, $p_k^{(\sigma)} = P_k z^{-k}$. Each chain of length $l$ contains $(l + 1 - k)$ 'k-mers', therefore

$$P_k = \sum_{l=k}^\infty (l + 1 - k) f_l$$

(12)

Note that $P_k$ has the maximum value of 1 which is approached in the limit when all chains are much longer than $k$.

We consider a problem of hybridization of polymers since the start of the night phase of the cycle when all of them are not hybridized. To describe the hybridization kinetics we use the fractions of fully hybridized k-mers $1 \geq \psi_k^{(\sigma)}(t) \geq 0$ as our dynamic variables. By definition, the concentration of such pairs of bound k-mers is $C \cdot p_k^{(\sigma)} \psi_k^{(\sigma)}(t)$. We note that hybridization states of different $k$-mers are not independent from each other since some of them overlap. To account for this, we introduce one more variable $\psi_k^{(\sigma)} \leq 1 - \phi_k^{(\sigma)}$ which is the fraction of all $k$-mers with a given sequence $\sigma$ that are available for hybridization. Now the binding kinetics of all $k$-mers can be described by the following set of coupled kinetic equations:

$$\tau \dot{\psi}_k^{(\sigma)} = C \cdot p_k^{(\sigma)} \psi_k^{(\sigma)} \psi_k^{(\sigma)} - \exp\left(\frac{\Delta G_{\sigma}}{k_B T}\right) \phi_k^{(\sigma)}$$

(13)

Here $1/\tau$ is the hybridization rate, $\Delta G_{\sigma}$ is the hybridization free energy for a given sequence $\sigma$, and $\sigma'$ is the sequence complementary to $\sigma$. For simplicity, we consider a symmetric case where mutually complementary $k$-mers have the same fraction, $p_k^{(\sigma)} = p_k^{(\sigma')}$. In order to solve these equation, one needs to specify a relationship between fraction of available k-mers $\psi_k^{(\sigma)}$ and hybridization probabilities, $\phi_k^{(\sigma)}$, that would take into account mutual overlap of the sequences. However, at
early stages the hybridization probability remains sufficiently low, and one can therefore assume
ψ_k^{(σ)} = ψ_k^{(σ')} ≈ 1 in Eq. (13). This results in a set of decoupled equations that can be analytically
solved (see SI). Thus calculated hybridization probability φ_k for different times Δt is shown in
Fig. 4.

Ligation-fragmentation kinetics The Eq. (4) describes the effective merger rate μ when lengths
n and m of two substrate chains hybridized to a template are longer than k_0. In a more general
case one needs to introduce length-dependent effective merger rate μ_{nm}. Under assumption of the
maximum sequence disorder this rate is given by:

\[ μ_{nm} = \lambda C^2 \sum_{k_1=1}^{\min(n,k_0)} \sum_{k_2=1}^{\min(m,k_0)} \frac{P_{k_1+k_2}}{2^{k_1+k_2}} \exp\left( -\frac{2\Delta G_0 + \Delta G_{st} + (k_1 + k_2) \cdot \Delta G}{k_B T} \right) = \lambda \left( \frac{C}{C_0} \right)^2 \cdot \sum_{k_1=1}^{\min(n,k_0)} \sum_{k_2=1}^{\min(m,k_0)} P_{k_1+k_2} \exp((k_1 + k_2) \cdot \epsilon) \] (14)

Here μ_{nm} corresponds to a particular order in which chains n and m merge into a longer chain.
Note that for directed chains such as nucleic acids there are two ways of merging chains, while for
undirected polymers there are four.

For directed polymers the resulting set of kinetic equations can be written as:

\[ \frac{1}{2\beta} \dot{f}_n = \left[ \frac{n}{2} + \Gamma^2 \sum_{m} \mu_{n,m} f_m \right] f_n + \Gamma^2 \sum_{m<n} \frac{(1 + \delta_{n-m,m})}{2} \mu_{m,n-m} f_m f_{n-m} + \sum_{m>n} f_m \] (15)

Here Γ is the dimensionless control parameter of the model which is proportional to monomer
density:

\[ Γ = \left( \frac{C}{C_0} \right) \sqrt{\frac{\lambda}{\beta}} \] (16)

22
and $\mu_{nm}$ is the "k-mer"- dependent ligation matrix:

$$
\mu_{nm} = \sum_{k_1=1}^{\min(n,k_0)} \sum_{k_2=1}^{\min(m,k_0)} P_{k_1+k_2} \exp (\epsilon \cdot (k_1 + k_2))
$$

This set of kinetic equations gives a complete description of the system in question and was numerically integrated to compare with our analytical results.

**References**

1. Wang, T. et al. Self-replication of information-bearing nanoscale patterns. *Nature* **478**, 225228 (2011).

2. Zeravcic, Z. and Brenner, M. P. Self-replicating colloidal clusters. *Proceedings of the National Academy of Sciences* **111**, 17481753 (2014).

3. Eigen, M. and Schuster, P. A principle of natural self-organization. *Naturwissenschaften* **64**, 541565 (1977).

4. Dyson, F. J. A model for the origin of life. *Journal of Molecular Evolution* **18**, 344350 (1982).

5. Kauffman, S. A. Autocatalytic sets of proteins. *Journal of theoretical biology* **119**, 124 (1986).

6. Jain, S. and Krishna, S. Autocatalytic sets and the growth of complexity in an evolutionary model. *Physical Review Letters* **81**, 56845687 (1998).

7. Doudna, J.A., Szostak, J.W. RNA-catalysed synthesis of complementary-strand RNA. *Nature* **339**, 519522 (1989).
8. Lincoln, T. A. and Joyce, G. F. Self-sustained replication of an RNA enzyme. *Science* **323**, 1229–1232 (2009).

9. Gilbert, W. Origin of life: The RNA world. *Nature* **319**, 618 (1986).

10. Orgel, L. E. Prebiotic chemistry and the origin of the RNA world. *Critical reviews in biochemistry and molecular biology* **39**, 99-123 (2004).

11. Robertson, M. P. and Joyce, G. F. The origins of the RNA world. *Cold Spring Harbor perspectives in biology* **4**, a003608 (2012).

12. Mast, C. B., Schink, S., Gerland, U. and Braun, D. Escalation of polymerization in a thermal gradient. *Proceedings of the National Academy of Sciences* **110**, 8030–8035 (2013).

13. Anderson, P. W. Suggested model for prebiotic evolution: The use of chaos. *Proceedings of the National Academy of Sciences* **80**, 3386–3390 (1983).

14. Anderson, P. W. and Stein, D. L. in Self-Organizing Systems (eds. Yates, F. E., Garfinkel, A., Walter, D. O. and Yates, G. B.) 445–457 (Springer US, 1987) based on 1981 preprint.

15. Derr, J. et al. Prebiotically plausible mechanisms increase compositional diversity of nucleic acid sequences. *Nucleic Acids Research* **40**, 4711–4722 (2012).

16. Bellini, T. et al. Liquid crystal self-assembly of random-sequence DNA oligomers. *Proceedings of the National Academy of Sciences* **109**, 11101–1115 (2012).

17. Rohatgi, R., Bartel, D.P., Szostak, J.W. Kinetic and mechanistic analysis of nonenzymatic, template-directed oligoribonucleotide ligation. *J Am Chem Soc* **118**, 3332–3339 (1996).
18. Vetsigian, K. and Goldenfeld, N. Genome rhetoric and the emergence of compositional bias. *Proceedings of the National Academy of Sciences* **106**, 215220 (2009).

**Acknowledgements**  Research was carried out in part at the Center for Functional Nanomaterials at Brookhaven National Laboratory, which is supported by the U.S. Department of Energy, Office of Basic Energy Sciences, under Contract No. DE-AC02-98CH10886. Work at Biosciences Department was supported by US Department of Energy Office of Biological Research Grant PM-031. We would like to thank Prof. Mark Lukin, Stony Brook University for valuable discussions.

**Author contributions**  A.T. and S.M designed the study and wrote the manuscript; A.T. performed the computational simulations;

**Correspondence**  Correspondence and requests for materials should be addressed to A.T. (email: oleksiyt@bnl.gov) and S.M. (email: maslov@bnl.gov)

**Competing Interests**  The authors declare that they have no competing financial interests.
Supplementary Information

Analytic results for hybridization kinetics As was described in the main text the set of kinetic equations for hybridization probabilities becomes decoupled in the regime when the probabilities are small \( \varphi_k(\sigma) \ll 1 \), and one can therefore assume \( \psi_k(\sigma) = \psi_k(\sigma') \approx 1 \). The Eq. (13) then becomes

\[
\tau \dot{\varphi}_k(\sigma) = C \cdot p_k(\sigma') - \exp \left( \frac{\Delta G_{\sigma}}{k_B T} \right) \varphi_k(\sigma) \quad (S1)
\]

The solution is exponential relaxation of hybridization variables \( \varphi_k(\sigma) \) towards their equilibrium values:

\[
\varphi_k(\sigma)(t) = K_k^*(\sigma) p_k(\sigma') \left( 1 - \exp \left( -\frac{t}{\tau_k(\sigma)} \right) \right) \quad (S2)
\]

In this expression,

\[
K_k^*(\sigma) = C \exp \left( -\frac{\Delta G_{\sigma}}{k_B T} \right) \quad (S3)
\]

\[
\tau_k(\sigma) = \tau \exp \left( -\frac{\Delta G_{\sigma}}{k_B T} \right) \quad . \quad (S4)
\]

The single most important factor that determines the hybridization free energy \( \Delta G_{\sigma} \) is the sequence length \( k \). For simplicity of the analysis we will replace \( K_k^*(\sigma) \) with its sequence- averaged value:

\[
K_k^*(\sigma) \approx K_k = C \exp \left( -\frac{\Delta G_0 + k \Delta G}{k_B T} \right) \quad (S5)
\]

This leads to the following result:

\[
\varphi_k(t) = C P_k z^{-k} \exp \left( -\frac{\Delta G_0 + k \Delta G}{k_B T} \right) \left( 1 - \exp \left[ -\frac{t}{\tau} \exp \left( \frac{\Delta G_0 + k \Delta G}{k_B T} \right) \right] \right) \quad (S6)
\]
As shown in Figure 4 at any given time \( t \) this expression is strongly peaked at a single value of \( k \), which weakly (logarithmically) depends on time:

\[
k \approx k_0(t) \simeq -\frac{k_B T}{\Delta G} \log \left( \frac{t}{\tau} \right) \tag{S7}
\]

\[
\varphi_{k_0} \simeq CP_{k_0} \exp \left( -\frac{\Delta G_0}{k_B T} + \varepsilon k_0 \right) \tag{S8}
\]

**Evaluating effects of finite \( \bar{L} \).** The equations (3) and (5) in the main text were derived in the limit \( \bar{L} \gg 1 \). Below we will relax these approximations to derive the exact formula working for arbitrary \( \bar{L} \).

In deriving the Eq. (3) in the main text we replaced the discrete summation with an integral. This approximation can be avoided by performing an explicit summation of the discrete geometric progression:

\[
\sum_{l=1}^{\infty} l \cdot \exp \left( -\frac{l}{\bar{L}} \right) = \exp \left( -\frac{1}{\bar{L}} \right) \frac{1}{\left[ 1 - \exp \left( -\frac{1}{\bar{L}} \right) \right]^2} = \frac{1}{4 \sinh \left( \frac{1}{2\bar{L}} \right)^2} .
\]

This amounts to replacing \( \bar{L} \) in Eq. (3) with \( \frac{1}{2 \sinh \left( \frac{1}{2\bar{L}} \right)} \):

\[
\frac{1}{2 \sinh \left( \frac{1}{2\bar{L}} \right)} = \sqrt{\frac{\mu}{\beta}} \tag{S9}
\]

The exact triple summation of the Eq. (4) in the main text

\[
\mu = \lambda \left( \frac{C}{C_0} \right)^2 \sum_{k_1=1}^{k_0} \exp(k_1 \cdot \varepsilon) \sum_{k_2=1}^{k_0} \exp(k_2 \cdot \varepsilon) \sum_{l=k_1+k_2}^{\infty} (l - k_1 - k_2 + 1) f_i
\]

for \( f_i \sim \exp(-l/\bar{L}) \) can be carried out in two steps. First, the sum over \( l \) combined with normalization \( \sum_l l \cdot f_i = 1 \) gives rise to

\[
\mu = \lambda \left( \frac{C}{C_0} \right)^2 \exp(1/\bar{L}) \sum_{k_1=1}^{k_0} \exp \left[ k_1 \cdot (\varepsilon - 1/\bar{L}) \right] \sum_{k_2=1}^{k_0} \exp \left[ k_2 \cdot (\varepsilon - 1/\bar{L}) \right]
\]
The discrete summation over \( k_1 \) and \( k_2 \) results in

\[
\mu = \lambda \left( \frac{C}{C_0} \right)^2 \exp(1/\bar{L}) \left( \frac{\exp[k_0(\epsilon - 1/\bar{L})] - 1}{1 - \exp(-\epsilon + 1/\bar{L})} \right)^2.
\]

(S10)

The Eq. (5) then becomes

\[
\frac{1}{2 \sinh \left( \frac{1}{2\bar{L}} \right)} \exp \left( \frac{k_0 - 1/2}{\bar{L}} \right) = \frac{C}{C_0} \cdot \sqrt{\frac{\lambda}{\beta}} \cdot \exp \left( \frac{k_0 \epsilon}{\bar{L}} \right) - \exp \left( \frac{k_0}{\bar{L}} \right) \cdot \frac{\exp(k_0 \epsilon) - \exp \left( \frac{k_0}{\bar{L}} \right)}{1 - \exp(-\epsilon + 1/L)}.
\]

(S11)

Here we neglected the exponentially small term in the enumerator of the r.h.s. of Eq. (S10). The green-dashed line in Fig. 3 shows \( \bar{L} \) defined by this equation plotted as a function of \( C \).