A mechanism for how prebiotic polymers may have become informational

Elizaveta Guseva,[*,†] Ronald N Zuckermann,[‡] and Ken A Dill[*,†]

Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY, (United States), and Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA (United States)

E-mail: elizaveta.guseva@stonybrook.edu; dill@laufercenter.org
Phone: +1631 632 5400. Fax: +1631 632 5405

Abstract

We propose a model for how the random prebiotic polymerization of short chains could have led to longer-chain autocatalytic sets of informational polymers. The central idea is to focus on polymers of hydrophobic (H) and polar (P) monomers, such as in present-day proteins. A principal tool for studying random sequences of such polymers is the HP lattice model. It is known that a significant fraction of random HP sequences can fold into relatively structured compact structures. We show that a significant fraction of those foldamers will have clusters of hydrophobic residues on the surface. For HP polymers, such binding sites can help catalyze the elongation of other HP polymers. A unique aspect of this proposal is that it gives specific guidance about what monomer sequences define the autocatalytic set, and a mechanism. Even in random "soups" of HP chains, enough of them may be able to fold and provide primitive catalytic surfaces to bootstrap the growth of the lengths and informational content of the set. We believe, this mechanism is relevant in the early origins of life.

Introduction

How might prebiotic polymerization processes have produced long chains of protein-like or nucleic-acid-like molecules[1,2]? Our aim here is to identify a particular physical process that could have operated during the early origins of life that was autocatalytic; that could have produced chains that are longer than are currently observed in prebiotic experiments; and that could have led to selective amplification of populations of particular sequences (on the road to biology) from underlying processes that are otherwise random chemical polymerizations.

We first give some background. The main question we ask here is how the Chemistry-To-Biology (CTB) transition might have happened to initiate the earliest life. Many chemical processes tend toward equilibria. Biological processes are non-equilibria that are both self-sustaining and self-supporting. Our focus is not on the chemical types of monomers and polymer, on on the conditions needed. Rather, our question here is the matter of physical principle of what chemical polymerization processes might lead from random to informational chains, from short to long chains, and to autocatalytic sets that could be self-sustaining. We review briefly below other work on those questions.

The Autocatalysis Puzzle

What molecules might have catalyzed the CTB transition? Early on, it was recognized by Eigen, Kaufmann, Dyson and others that the

[*] To whom correspondence should be addressed

[†] Stony Brook University

[‡] Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA (United States)

CTB transition requires autocatalysis, ie some form of positive feedback or bootstrapping.[3–6] These works established the importance of autocatalysis for sustaining stable biochemical networks. Since then there were many works, which studied various artificial chemical systems with different implementations of positive feedback. Various aspects of these systems has been investigated, such as for example ability to reach biologically reasonable lengths (for polymeric systems), maintain exponential growth, inheritance, evolvability. And what mechanism of positive feedback gives all the desirable properties if any. Here we provide several example of such works.

In several works,[7–9] an artificial chemistry system is modeled as sustained by mutually catalyzing small molecules. Collections of small molecules of different types, some of which were randomly chosen to be chemically active, were allowed to interact and their evolution through time and cell division has been studied. Those works showed that such systems can maintain a collective identity, expressed through the types of molecules (the "composome") that survive through cell "divisions", and respond to certain changes of the external conditions. This model corresponds to a lipid world scenario, but being a rather general metabolism-first model, it can be used to study various aspects of systems of mutually interacting small molecules, which is a very interesting area of studies given the attractiveness of the metabolism-first scenario and problems in encounters in close-to-biological chemistries (see for example[10]).

Wu and Higgs[11] have shown how autocatalysis in mono-polymeric artificial chemistry systems produce a bi-stability and increase chain lengths of polymers. The model is very simple, but incorporates only one type of monomers, which removes evolution, diversity and effects of "parasitic" species on a system out of consideration and makes it a very simplistic toy model. A similar theoretical model shows how long-chain polymers can be self-sustaining by template-assisted ligation and random breakage,[12] they show that under certain condition, system goes through a phase transition, where system produces longer chains and maintain it self. The system effectively doesn't distinguish between polymers of different composition and therefore is the subject to similar limitations as.[11]

Theoretical studies of binary polymers either capable of autocatalysis or replication.[13–16] The system used in these studies is a very simple one, with unlimited "food" molecules, which restricts the model only to non-competing systems. The autocatalysis mechanism of these series is very simple one: a selected sequence has an ability to accelerate growth of it's own precursors only (either full or partial). This limits results of the models to a RNA-like-world, with molecules capable of direct templating (in the case of replication) or of some sort of self-recognition (in the case of catalysis),maybe also through direct templating. The studies showed that while autocatalytic system has bi-stability and increased ratio of longer polymers, one has to increase catalysis rate exponentially in order to get exponential growth of longer chains. Self-replicating systems on the other hand didn't show bi-stability, but substantial polymer growth was found to arise from low replication rates. It was also shown that self-replication enhances diversity of the system and compensate for possible chemical bias. The nature of the model, may however severely limit its application, due to the fact that it focuses only on molecules already capable of self-recognition, while recognition of "self" is in fact a difficult question in chemistry of the origin of life.

Significant progress has been made in attempts to make artificial autocatalytic sets in the laboratory.[17–19] Such systems are designed so that a pair of molecules catalyze each other, giving autocatalysis and exponential growth. These studies are important for our understanding of autocatalysis, but are simple toy models. [E - Toy models how? What important aspects are they missing?]

The 'Flory Problem': polymerization processes produce mostly short chains

A key puzzle in biopolymer origins is how to produce sufficiently long chains? Many experiments show that either amino acids or nucleotides can polymerize into short-chain molecules under prebiotic conditions without the presence of enzymes.[20–24] It is also known that the yields of such short-chain oligomers can be increased under prebiotically plausible conditions by such processes as adsorption to clays[25,26] or minerals,[27,28] by evaporation of tidal pools,[29] by concentration in ice through eutectic melts[30] or freezing[31] or temperature cycles.

It remains a puzzle, however, to understand how prebiotic processes could have overcome what we call the "Flory Problem" – the production of long-chain polymers. It is generally assumed that the minimum chain lengths of proteins or nucleic acids that are needed to make the complex structures essential for biological function is estimated to be around 30-60 monomers long.[32] Yet, since the earliest work of Flory and others, which elucidated basic mechanisms of chain polymerizations, it has been clear that chain syntheses are stochastic, whereby the concentrations of longer chains are exponentially smaller than of shorter chains.

Correspondingly, experimental studies which are intended to find out possible ways of prebiotic polymerizations of amino acids and nucleotides lead predominantly to short chains. Leman et al. showed that carbonyl sulfide (COS), a simple volcanic gas, brings about the formation of oligo-peptides from amino acids under mild conditions in aqueous solution in minutes to hours. But the product is mainly dimers and trimers.[23] In another study, using various mineral catalysts such as calcium montmorillonite, hectorite, silica or alumina, mixtures of Gly and Gly_2 grow to about 6-mers after 14 days.[33,34] Or, by freezing samples of phosphoimidazolide-activated uridine in the presence of metal ions in dilute solutions, Kanavarioti found polymers of oligouridylates up to 11 bases long, with an average length of 4.[30] And, starting from decanucleotides $[^{32}P]dA(pdA)_8pA$ adsorbed on Na^+-montmorillonite, Ferris et al. observed chains averaging lengths 20-40 after 14 days at 25℃.[28] It is not yet understood how prebiotic polymerizations could lead to the types of long protein or nucleic-acid chains that are found in present-day cells.

The standard stochastic processes of chain polymerization lead to the the Flory or Flory-Schulz distribution of the concentrations of chains of different chain lengths.[35]

$$f(a) = a^2 l (1-a)^{l-1}, \tag{1}$$

where l is the chain length and a is the probability of chain termination, which is a measure of the average chain length: $\langle l \rangle = a(2-a)$. Figure 1 shows the central prediction of Flory theory, that longer chains are exponentially less populated than shorter chains. For example (see the blue line in Fig 1):

$$\frac{[10 \text{ mers}]}{[1 \text{ mers}]} \propto 10^{-4}, \qquad \frac{[20 \text{ mers}]}{[1 \text{ mers}]} \propto 10^{-9} \tag{2}$$

Thus, for a synthetic process that starts with micro-molar concentrations of monomers, the average chain length would be $\langle l \rangle = 2$ and 40-mers would have negligible concentrations of $\propto 10^{-19}$ mol/L.

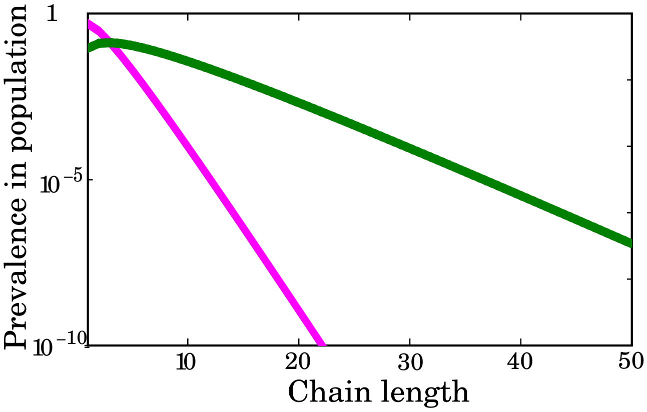

Figure 1: The Flory length distribution that arises from spontaneous polymerization processes. Green line gives $\bar{l} = 6$, magenta one corresponds to $\bar{l} = 2$

The Flory distribution is a good model for various known prebiotic syntheses of peptide and nucleic acid chains. Figure 2 shows that the Flory model fits known length distributions from various prebiotic syntheses (sometimes data is fit to an exponential law, $f(a) \propto const^{l}$, which is slightly simpler but nearly identical in form [13,16]). So, since experiments on a.a. polymerization and prebiotic chemistry appears to give submillimolar or submicromolar concentrations[30,36–39] of monomers, longer chains have been expected to be present in negligible concentrations. The Flory problem is not solved by improving the catalyst.[16] The problem is in the equilibrium, not the kinetics. The slope of the Flory plot is governed by a, the equilibrium constant for bonding each monomer into the chain.

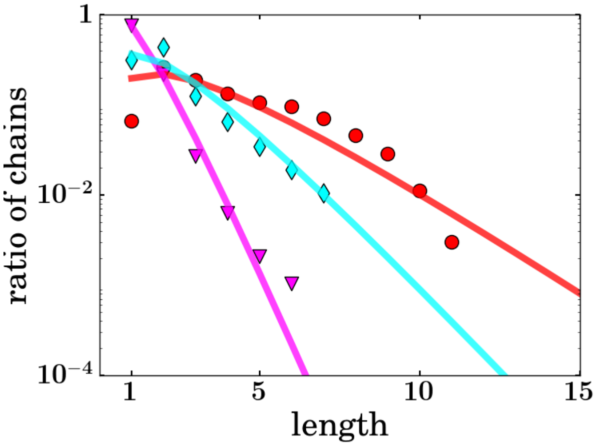

Figure 2: Chain-length distributions of peptides and nucleic acids synthesized under prebiotic conditions. The lines show curve fits to the Flory distribution for polymerization processes. Data points: red – Kanavarioti,[30] cyan – Ding,[40] magenta – Ferris[41]

Proposed mechanism: Short HP chains fold in water and catalyze the elongation of other HP chains

We believe that some puzzles of prebiotic chemistry might be resolved by recognizing the special properties of HP polymers. HP polymers are copolymers in which the monomer units can be categorized as hydrophobic (H) or polar (P), in particular sequences. Proteins are present-day HP polymers: the 20 amino acids can be divided into the two classes, H or P. In HP polymers, the sequence patterning of the H and P monomers lead to a solvation-based encoding of sequence-structure relationships.[42] The 2D HP model is the model of choice for studying folding-related properties of random sequences, because the full sequence and conformational spaces can be studied by exhaustive enumeration, without assumption or adjustable parameters.

Here is the overview of the mechanism that is supported by our model simulations below. Suppose that a polymer molecule A folds so that it has exposed hydrophobic monomers on its surface. This patch can serve as a sticky spot for another HP molecule B and for an H monomer C. Hydrophobic interaction between three of them localizes growing chain and next monomer and reduces the kinetic barrier of polymerization for B and C molecules; see Fig. 3. A typical hydrophobic interaction is $1 - 2kT$. Consequently chain A is a catalyst that provides a hydrophobic landing pad and reduces activation energy by 3-4 hydrophobic interactions, thereby increasing the polymerization rate around 100-fold (fig. **??**). Of course, this rate enhancement is much smaller than the 10^{7}-fold of modern ribosomes.[43] Nevertheless, it gives a conceptual basis for how peptide-bond formation rate enhancements might have had their prebiotic beginnings. Our studies below lead to two main conclusions. First, even random processes that synthesize random HP sequences will lead to some selection that can concentrate some sequences and structures over others. Second, some folded HP sequences can have primitive catalytic and autocatalytic abilities, based on their exposed hydrophobic surfaces, whereby certain chains can help to polymerize other chains. So, the folding and binding sites of HP polymers could explain how prebiotic chemistry might escape the Flory Problem.

HP polymers have been studied extensively as a model for the folding and evolution of proteins.[42,44–47] Such studies have shown that sta-

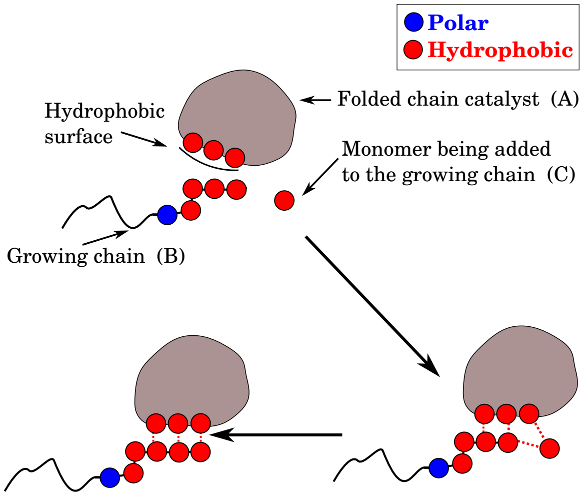
● Polar
● Hydrophobic
Hydrophobic surface
Folded chain catalyst (A)
Monomer being added to the growing chain (C)
Growing chain (B)

Figure 3: Chain A folds so that it has exposed hydrophobic monomers on its surface. This patch can serve as a sticky spot for another HP molecule B and for an H monomer C. Hydrophobic interaction between three of them localizes growing chain and next monomer and reduces the kinetic barrier of polymerization for B and C molecules by 3-4 hydrophobic interactions, increasing the polymerization rate around 100-fold

bly folded structures of proteins to a large extent can be explained by binary pattern of polar and hydrophobic residues and do not require knowledge of specific interresidue contacts.[48–50] A large fraction of the space of random sequences can collapse into compact structures resembling native proteins;[44] see fig. 4. While the 2-dimensional HP lattice model entails obvious simplifications, it has the advantages that: (1) it is currently the only model that can explore full sequence and conformational spaces, without assumptions or approximations, and (2) it is known to reproduce key observations on real proteins in 3D. The reason that the dimensionality is not problematic is that the principle physics is in the surface-to-volume ratios of HP chains collapsing in water, and the 2D model for 12-25-mers in 2D is the same as that of 100-200-mer proteins in 3D.[51]

We note that the present model is not particularly intended to be restricted to proteins. RNA molecules are also able to fold in water, indicating differential solvation. While our present model focuses on hydrophobic interactions, it is simply intended as a concrete model of solvation, that could more broadly include hydrogen bonding or other interactions. So, while our analysis below is only applicable to foldamers, it is not limited to proteins.

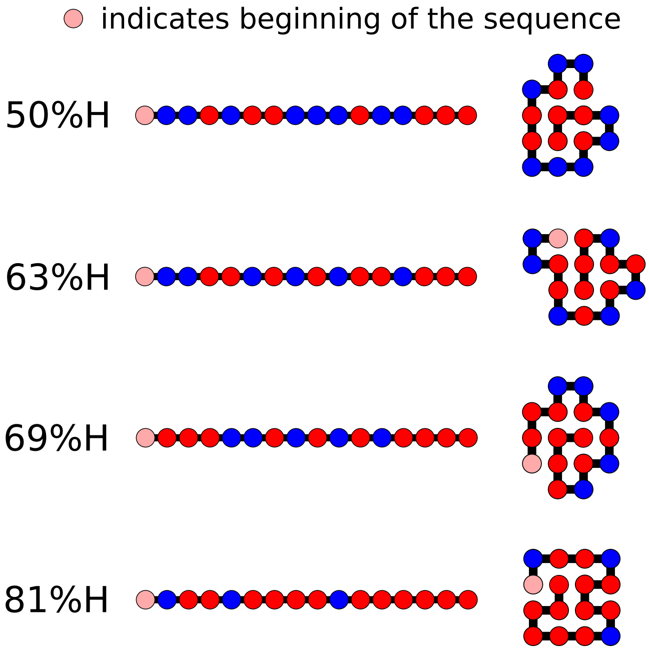

Figure 4: HH interactions are favorable in water, leading to different compact states for different HP sequences.

The model of HP polymerization kinetics

We consider a soup containing a sufficient supply of activated H and P monomers in solution. Activated H and P monomers are supplied by an external source with rate a. Any molecule can be removed from the system with rate d. A given chain elongates through the an addition of a monomer, at rate α. Without loss of generality we define the unit rate by setting $\alpha = 1$. All other rates are taken relative to this chain growth rate. We assume all polymers also undergo spontaneous hydrolysis; any bond can be broken with rate constant h.

Now, in addition to this basic polymerization dynamics, we also account for the fact that some sequences will fold into native HP structures, protecting its hydrophobic core residues from hydrolysis. In the standard definition for HP polymers, we regard a chain as folded if it has a unique lowest-energy structure. That is, some sequences have only a single conformation giving a maximum number of HH non-covalent contacts. (Most sequences, in contrast, have many low-energy states; we do not count these as folded structures.) The conformational energy of the native fold (E_{nat}) of any particular folded sequence equals the number of hydrophobic interactions ($n_{h\phi}$) × the energy E_H, of one hydrophobic interaction, which is known to be $\approx 1 - 2kT$:[52]

$$E_{nat} = n_{h\phi} E_H. \tag{3}$$

This energy differs for different sequences. Now, given knowledge of E_{nat} for any particular sequence, we can readily compute the folding and unfolding rate coefficients from:[52]

$$\ln\left(\frac{k_f}{k_u}\right) = -\Delta G/kT = E_{nat}/kT - N\ln z, \tag{4}$$

for reversible folding, where z is the number of rotational degrees of freedom per peptide bond.

We suppose that chains that fold are prevented from further growth, and also are protected from hydrolysis. This simply reflects that open chains are much more accessible to degradation from the solvent or adsorption onto surfaces than are folded chains. Even so, folding in our model is a reversible equilibrium, as it is for natural proteins, so some small fraction of the time even folded chains are unfolded, and in that proportion, our model allows them further growth and degradation.

Results: some basic tests of the dynamics of polymerization and folding

We performed a computational experiment just to test that this dynamical model leads to the Flory equilibrium distribution. In this simple test, we excluded folding and catalysis by setting the hydrophobic energy to 0. Figure 5(a) shows that, as expected, the length distribution has an exponential tail. Figure 5(b) shows populations of individual sequences in a randomly chosen realisation of the experiment. For more details see Simulations. Experiment 1.

Folding alone does not solve the Flory Problem

Our second computational experiment asks whether HP chain folding alone is sufficient to alter the Flory length distribution. In short, we ask whether the few randomly synthesized sequences that happen to fold, which therefore also happen to protect their interior residues from hydrolysis, lead to changing the Flory length distribution. Details are in experiment 2; see Simulations. Figure 6 shows that chain folding alone, of the few foldable sequences, is not sufficient to escape the Flory problem of the exponentially diminishing concentrations of chains with length. Folding does increase the abundances of some of the folded sequences. Yet, their instances are so rare that they do not alter the distribution. Thus, folding alone does not explain how prebiotic polymers escape the Flory Problem.

Primitive foldamers can also be primitive catalysts

In our third experiment, we also accounted for the effect of some HP foldamers to catalyze reactions. Of course, present-day foldamers (proteins and RNA molecules) can catalyze reactions very efficiently – including the reaction in the ribosome that synthesizes the peptide bond.[53] These modern catalysts have evolved over a long time towards exceptional catalytic functionalities. Of course, prebiotic functionalities are likely to have been much more primitive. Our interest here is in how the most primitive catalysts might have begun. Precision and complexity isn't a requirement for peptides to perform biological function. Studies show that folded proteins generated from random libraries can sustain the growth of living cells[50] and specific bindings between them and small molecules are not rare.[54] These findings imply that some primitive foldamers could have had some primitive catalytic capability. The unique power in catalysis of a foldamer – in contrast to other polymeric structures – is that it can

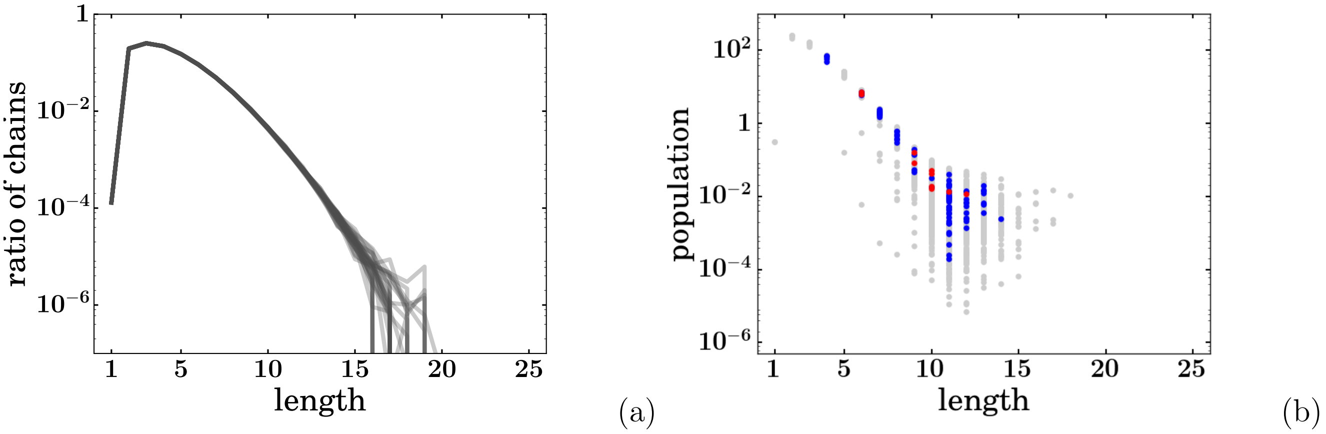

Figure 5: Results of the Experiment 1: no folding, no catalysis polymerization. This is the base case to which we compare all other systems. Each data point is an average of 10^6 time points in the steady state interval. (a) A single line shows length distribution for one simulation run (we run total of 30 simulations). Populations decay exponentially as length grows. (b) Populations of the individual sequences, based on the results of a single simulation run. We set $E_h = 0$, but still show sequences,which could've fold (if not $E_h = 0$) as blue, and sequences which could've catalyze as red. As expected, they follow the same distribution as other sequences.

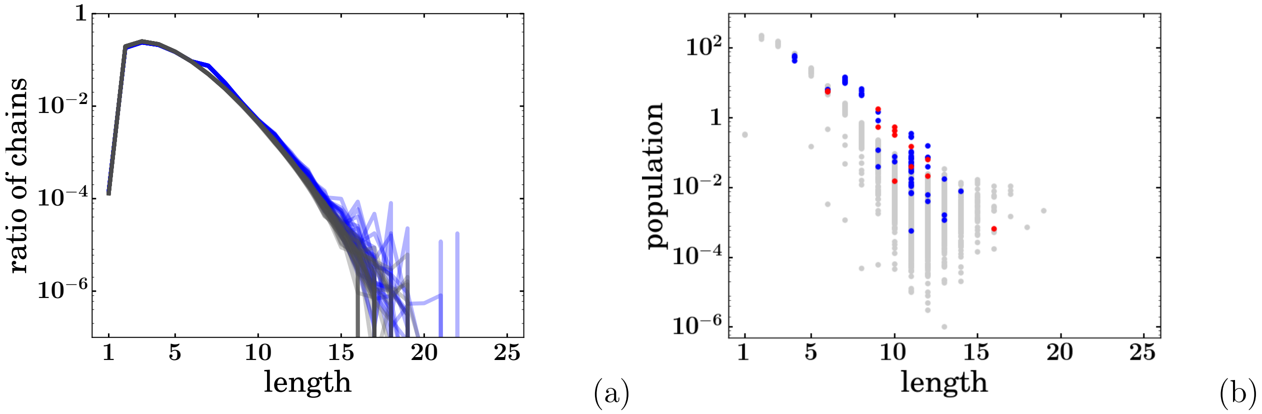

Figure 6: Results of the Experiment 2: Polymerization with folding, but with no catalysis. Each data point is an average of 10^6 time points in the steady state interval. (a) A single line shows length distribution for one simulation run (we run total of 30 simulations). Blue lines are result of the Experiment 2, gray ones – Experiment 1. Addition of folding to the basic model doesn't change the nature of distribution. (b) Populations of the individual sequences, based on the relults of a single simulation run. Gray dots – unfoldable sequences, blue ones – foldable. In this simulation we disallow catalysis, but we still show potential catalysts in red, to show, that they behave similarly to regular folders. Foldable sequence have advantage compared to unfoldable ones.

resemble a microscale solid, with very precise positioning of different chemical moieties over sufficiently long time scales that substrates and transition states can 'recognize', bind, and react on them. For example, serine proteases utilize a catalytic triad of 3 amino acids. So, foldability in some type of prebiotic polymer, could conceivably have had a special role in allowing for primitive catalysis. Here, we use a toy model to capture that simple idea, namely that a folded polymer can position a small number of residues in a way that can catalyze a reaction.

How chain elongation is catalyzed

For those sequences same force that forces them inside would attract another hydrophobes from the solution If the exposed patch is long enough, then it can serve as a landing pad for a growing sequence and hydrophobic monomer, thus holding both together and lowering the activation barrier for the polymerization reaction.

Therefore catalysis rate is proportional to the exponent of hydrophobic energy E_H and number of contacting hydrophobes n_c: $\alpha \cdot \exp(E_H \cdot n_c/kT)$(see figure 7). For the modeling we set minimum size of the landing pad to be 3. Thus catalysts facilitate $\cdots HH$ to H connections. So every polymer, which has HHH in its sequence can benefit from catalysis.

At this point, we note what our model is, and what it is not. Our model is not intended as an accurate atomistic depiction of a real catalytic mechanism. It is a coarse-grained toy model, of which there will be variants. The mechanism we explore here is the translational localization of the two reactants, polymer B and monomer C, in the chain extension reaction. And, while this model is 2-dimensional, extensive previous studies have shown that it captures many important principles of folding and sequence-to-structure relationships. At the present time, this type of model is the only unbiased, complete and practical way to explore plausibilities of physical hyotheses such as the present one.

Modeling shows that HP foldamer-catalysts can solve the Flory Problem

Figure 6 shows the results of simulations, now allowing for both the folding of all HP sequences, according to the rules of the HP model, and allowing for catalysis based on all those foldamers that have three H monomers on their surfaces in their unique native folded states (see SI for details). Presence of catalysis in the system skews the distribution significantly. This distribution is fairly stable towards hydrolysis and dilution parameters. It allows for 1 order of magnitude change in those parameters without significant change in the behavior of the system.

This figure represent one of the key finding of this paper: a simple physical mechanism, such as hydrophobic interaction is capable of generating autocatalytic sets of non-trivial structure (for details see in Discussion), which despite the initial weak catalytic forces of interaction ($E_h \approx 1 - 2kT$) over time reaches the ability to dominate population. As it's clear from the figure 9(a) this mechanism also allow for solution of the Flory problem.

Figure 9(b) show populations of individual sequence as function of their length. Colors represent type of the sequence: red – for autocatalysts, blue – for foldable, gray – for the rest. While autocatalysts constitute a minority of the sequence space ($\approx 0.6\%$ of all the represented in the experiment sequences; foldable but inactive sequences occupy $\approx 2.3\%$, regular sequences – $\approx 97.1\%$), their contribution to the total mass is impressive: $\approx 15.7\%$; inactive folders – $\approx 29.2\%$; regular sequences $\approx 55.1\%$. Moreover, the longer the chain length, the higher the input of autocatalysts into total mass of that length (see figure 10). This happens first of all due to increasing number of autocatalysts among longer sequences (see fig.8 and also due to the fact that folding along isn't capable of reaching longer chains. While at the shorter chains preservation from hydrolysis by means of folding is enough at longer chains active influence of catalysis is necessary.

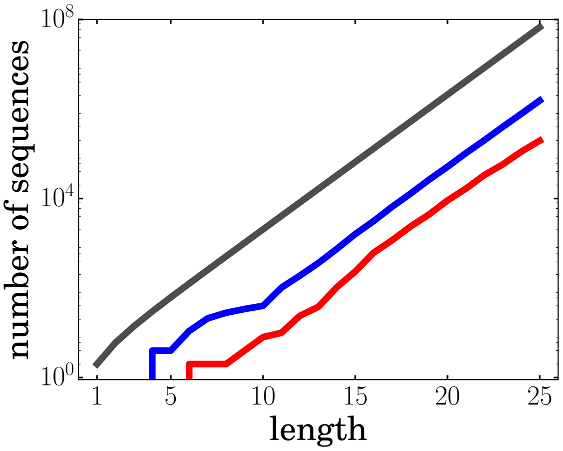

Figure 8: Number of all sequences (gray), foldable (blue) and catalytic (red) up to length x ,

Discussion

It has long been recognized that life's origins require some form of autocatalysis.[3,4,6] But, what mechanism at the molecular structural level might explain some type of molecular structural bootstrapping? Here, we find that a form

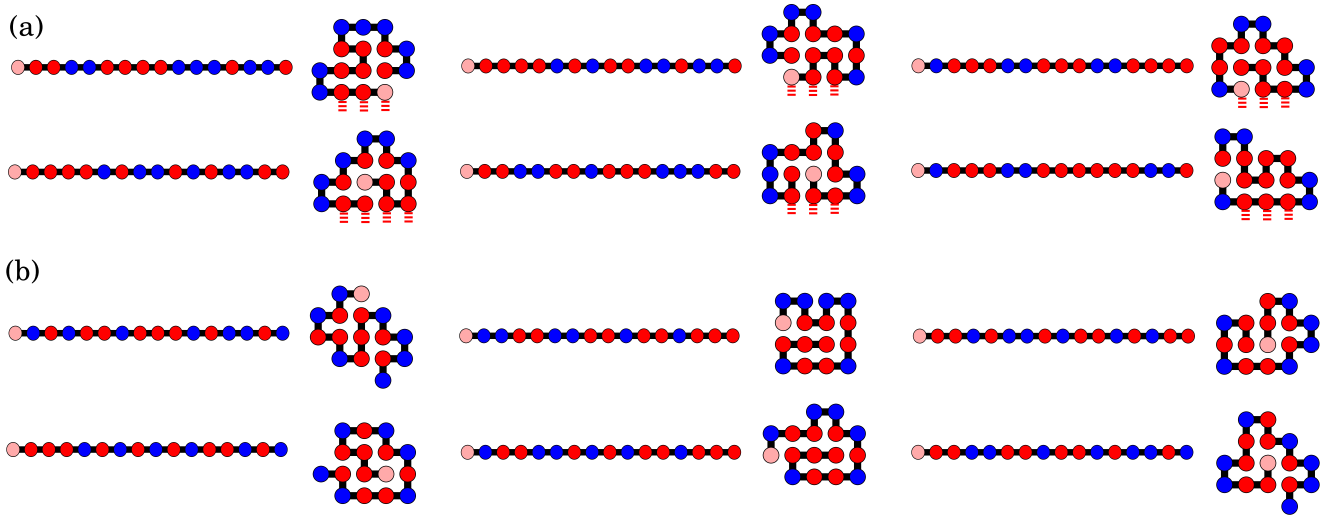

Figure 7: (a) When folded chain has a patch of at least three exposed nonpolar monomers, it can be a catalyst. This patch serves as a landing site for the growing chain and a hydrophobic monomer, facilitating the action of linking the monomer to the chain, by holding them together and lowering activation energy. (b) Most chains in their native states don't have enough hydrophobes exposed and cannot serve as catalysts.

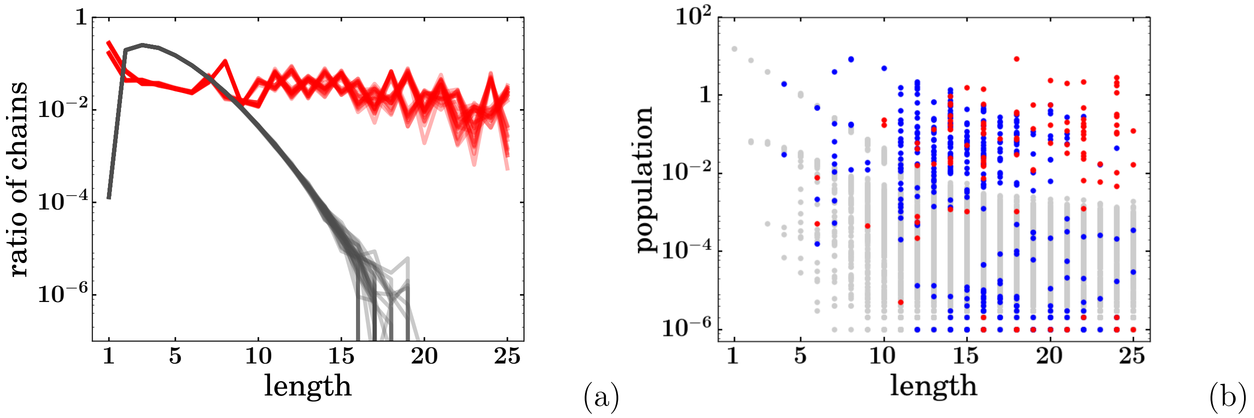

Figure 9: Results of the Experiment 3: polymerization with folding and catalysis enabled. Each data point is an average of 10^6 time points in the steady state interval. (a) Gray lines represent polymerization without folding or catalysis. Red ones corresponds to a simulation run with folding and catalysis. A single line shows length distribution for one simulation run (we run total of 30 simulations). For details of simulations see section Simulations, Experiment 3. (b) Populations of individual sequences are shown as functions of their length. Autocatalytic sequences are shown in red, sequences that can fold but cannot act as catalysts – in blue, and all the other sequences in gray. Lower limit of 10^{-6} is due to computational precision – there are 10^6 time steps over which we calculate average to get a point on the graph. Therefore the minimum possible population correspond to the case when sequence appeared only for one time instance. The data is an example based on one simulation

of autocatalysis, or positive feedback, is inherent in the following process: HP polymers are synthesized randomly; a small fraction of those HP polymers fold into relatively stable compact states; a fraction of those folded structures provide relatively stable 'landing pad' hydrophobic surfaces; those surfaces can help to catalyze the elongation of other HP molecules having foldable sequences. Figure 11 illustrates this process.

The HP model allows for precise counting of sequences that fold, or don't fold, and have any particular structural property. A non-negligible fraction of all possible HP sequences fold to unique structures (2.3% for lengths up to 25-mers). The fraction of all possible HP sequences that have catalytic surfaces (as defined above) is 12.7% of foldable sequences, or 0.3% of the

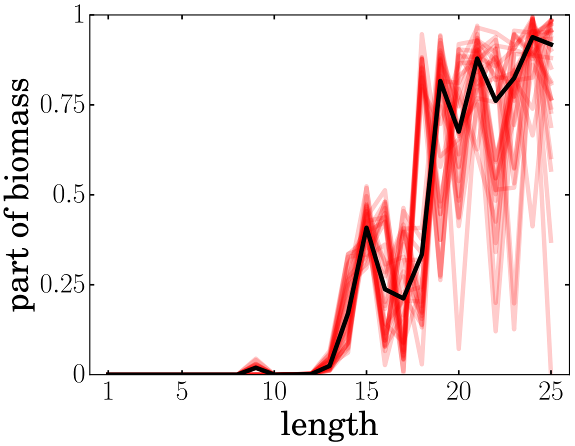

Figure 10: A single red line shows what ratio of the biomass of the given length is due to autocatalysts for a single simulation run. Each data point is an average of 10^6 time points in the steady state interval. Black line is a median over 30 simulations

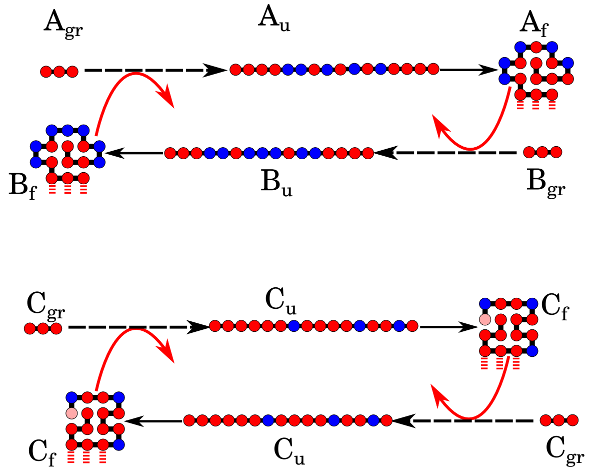

Figure 11: Two examples of cross and autocatalytic interaction naturally embedded in HP-polymers dynamics. Dashed arrows (- - -) are shorten representation of multiple reactions of chain among which there are $\cdots HH{-}H$ catalyzed reactions. Catalysis is represented by red solid arrows (—). Solid black lines (—) are folding reactions. Chains, which we call "autocatalytic" experience catalysis during one (or more often several) of the steps of elongation. Then, when they reach the length at which they can fold (A_u, B_u, C_u), they fold and serve as catalysts them selves (A_f, B_f, C_f). Mutual catalysis cat happen between different sequences (here A and B) and between different instances of the same sequence (here C).

whole sequence space. These ratios remain relatively constant with chain length, at least up to 25-mers; see figure 8. This and successful designs of foldable, biologically active proteins based on the HP folding rule[55] suggests that folding in HP polymers is not rare.

The present model provides an experimentally testable prediction for what early polymer sequences could be autocatalytic, and provides a structural and kinetic mechanism for their action. This model also provides a view about how selection and diversity may be related, and may be evolvable, arising from chain syntheses that are otherwise random.

Materials and methods

Simulations

To test our hypothesis we performed direct stochastic simulations on several sets of parameters. We used the PDMmod method[?] [1]. Stochastic simulations keep track of each molecular specie in the system. However simulations are limited due to computational reasons. First of all we have to explore conformational space of every polymer. This task is NP-hard (we use HPSandbox algorithm[44,56] [2]), so we had to limit maximum chain lengths to 25. We also try to keep total number of species in the low thousands, to limit computational costs. We do it by introducing dilution parameter d: molecules are being removed from the system with probabilities $\propto d$. This either can mimic a protocell splitting and loose of materials due to it or in the case when system isn't bounded by any borders the fact that some molecules will diffuse away. Total number of molecules varies from simulation to simulation, however it mostly holds in the region $10^2 - 10^4$.

We start our simulations with a small pool of monomers, usually below 100 molecules.

- Monomers can be react with each other and with polymers to produce polymerization reaction with rate $\alpha = 1$

$$\text{1-mer} + n\text{-mer} \xrightarrow{\alpha} (n+1)\text{mer} \qquad (5)$$

[1] C++ library and description can be found here: https://github.com/abernatskiy/pdmmod

[2] Python implementation and description can be found here: http://hp-lattice.readthedocs.org/en/latest/

- These monomers are being imported in the system. with rate $a \gg 1$. It is safe to assume that we would have enough monomers in the system and import of monomers wouldn't be a bottleneck of reactions chain. Therefore we explore big values of $a \propto 10^3\alpha$

$$\emptyset \xrightarrow{a} H \text{ or } P \tag{6}$$

- Hydrolysis has constant rate h per bond. Half-life time of hydrolysis bonds in neutral conditions and temperatures around room temperature are on the order of hundreds of years[3]. We test hydrolysis rate constants to be about $0.01 - 1$ of polymerization rate constants. This way we account for polymerization conditions, which happens on the order of days to years.

$$n\text{-mer} \xrightarrow{h} l\text{-mer} + (n-l)\text{-mer} \tag{7}$$

- Dilution parameter d mimics cell division and loss of the matter because of that. This parameter also serves utilitarian role of limiting total population of the cell. We explore valued of d from $\propto 0.01\alpha$ to $\propto 1\alpha$. Given values of a we'll explore various populations from $\propto 10^2$ to $\propto 10^4$ polymers per cell.

$$\text{anything} \xrightarrow{d} \emptyset \tag{8}$$

- Folding and unfolding reactions happen very quickly with the unfolding rate constants of $k_f \gg k_u \gg \alpha$.

$$\begin{aligned} \text{folded chain} &\xrightarrow{k_u} \text{unfoled chain} \\ \text{unfoled chain} &\xrightarrow{k_f} \text{folded chain} \end{aligned} \tag{9}$$

  The folding rate constant taken from:[52]

$$\ln\left(\frac{k_f}{k_u}\right) = -\Delta G/kT = E_{nat}/kT - N\ln z, \tag{10}$$

  where z is the number of rotational degrees of freedom per peptide bond. Explain calculations For proteins in aqueous solution, the parameters are known,[59,60] giving the following expressions for folding and unfolding rates:

$$\begin{aligned} &k_u = \exp[12 - 0.1\sqrt{N} - E_H(0.5N + 1.34)], \\ &k_f = k_u \exp(\Delta G) \end{aligned} \tag{11}$$

  E_h in our experiments is around $1-2$kT[?] . k_{unf} we keep $\propto 10^2$, which gives us range of unfolding rates from a reaction per hours and days and range of folding rates from a reaction per hours to fractions of a second.

- Catalysis rate is proportional to the exponent of hydrophobic energy E_h and number of contacting hydrophobes n_c: $\alpha_{cat} = \alpha \cdot \exp(E_h \cdot n_c/kT)$. Number of hydrophobic contacts for the short HP-sequences varies in the range $3-6$.Ther reaction is:

$$\text{Cat.} + H + \underbrace{\cdots HH}_{l-1} \xrightarrow{\alpha_{cat}} \text{Cat.} + \underbrace{\cdots HHH} \tag{12}$$

  With the hydrophobic energies of $1-2$kT this gives us catalysis rates around hours and days for one reaction. Because the PDMmod supports only binary reactions, we divided the reaction above into to steps: interaction of catalyst with a monomer with rate α and reaction of this complex with a polymer with the rate α_{cat}.

We investigated the behavior of the system in the steady state. To determined, when it was reached we looked at the total populations of all the chains, of all folders and of all catalysts separately. When all the populations stopped growing and started fluctuating around constant value, we say the steady state was

[3] Hydrolysis rate constants of oligopeptides in neutral conditions are of the order of $10^{-11} - 10^{-10}$: $1.310^{-10}M^{-1}s^{-1}$ for benzoylglycylphenylalanine ($t_{1/2} = 128y$),[57] $6.310^{-11}M^{-1}s^{-1}$ ($t_{1/2} = 350y$) for glycylglycine and $9.310^{-11}M^{-1}s^{-1}$ for glycylvaline.[58]

reached. In order to account for stochasticity we repeated every simulation for 30 times for every experiment, and considered the resulting ensembles. We ran all the simulation for $140s$ of internal simulation time during which $10^6 - 10^9$ individual reactions has occurred. We took measurements every $10^{-6}s$. For all the trajectories steady state behavior was reached no later than $40s$ from the start of a simulation. Thus we considered only last $100s$ (one million recordings) for each simulation. All the data points we used in the figures are averages over these recordings.

For all the experiments below we have the following parameters:

1. $\alpha = 1$
2. $a = 1000$

   With values $a \ll 1000$ or $a \gg 1000$ some of the experiments have total number of sequences and populations either to high to calculate or to low to make conclusions. Values around $a = 1000$ allow to run all the experiments with the same parameters without those complications.
3. $h = d = 0.1$.

   When $3d \lessapprox h \leq \alpha$, hydrolysis is very strong and in non-catalytic case there's explosion of short sequences, which makes simulations computationally nearly impossible.

   When $3h \lessapprox d \leq \alpha$, hydrolysis is very weak and nothing limits the growth of longer sequences, and with the chains shorter than 25-mers, there are considerable populations of 10-20-mers even without any folding or catalysis, besides high d makes total populations too low for any statistical calculations.

   When $0.05 \lessapprox d \approx h \lessapprox 0.5$ the forces of dilution and hydrolysis are relatively balanced and populations don't drop and don't explode.
4. $E_h = 2kT$
5. $z = 1.2$

The simulations were performed on the Laufer Center's computing cluster of CPUs. Source files of the models, parameters, initial conditions and random seeds can be obtained at `https://github.com/gelisa/hp_world_data`

Experiment 1. Reproduction of Flory distribution. We started simulations with a small pool of chains up to 3-mers. To calculate length distribution, for each trajectory we calculated average population of every sequence over time over all recordings after 40s, resulting in a million time steps, then we sum up all the populations of a given length, get total populations for all n-mers, $n \in [1, 25]$, and then divide every population to the sum of them:

$$p_n = \frac{\sum \text{all n-mers}}{\sum \text{total population}} \tag{13}$$

giving probability to encounter an n-mer, when randomly taking a chain out of a "tube"

The source file of the model and parameters of the simulation are located at `https://github.com/gelisa/hp_world_data/tree/master/001`

Experiment 2. Introduction of HP-folding We start with the same starting population as in Experiment 1. But now we introduce hydrophobic energy $E_h = 2kT$. To calculate length distribution, for each trajectory we calculated average population of every sequence over time over all recordings after 40s, resulting in a million time steps. The source file of the model and parameters of the simulation are located at `https://github.com/gelisa/hp_world_data/tree/master/002`

Experiment 3. Introduction of HP-catalysis. In addition to folding in this *in-silico* experiment we introduced interaction between proteins. All parameters are as above. We varied parameters of the simulations, and noticed significant stability of the length distribution towards change of h and d: $0.05 \lessapprox d \approx h \lessapprox 0.5$. Distribution is very sensitive towards hydrophobic energy, as expected. Chain length distribution is drastically different compared to Experiments 1 and 2 in the region when $E_h = 1 - 3kT$

References

(1) Joyce, G. *Cold Spring Harbor Symposia on Quantitative Biology* **1987**, *52*, 41–51.

(2) Abel, D. L.; Trevors, J. T. *Theoretical Biology and Medical Modelling* **2005**, *2*, 29.

(3) Eigen, M.; Schuster, P. *Naturwissenschaften* **1978**, *65*, 7–41.

(4) Dyson, F. *Origins of Life*; Cambridge: University Press, 1985.

(5) Prigogine, I.; Nicolis, G. *Exploring Complexity*; 1989.

(6) Kauffman, S. A. *Journal of Theoretical Biology* **1986**, *119*, 1–24.

(7) Segré, D.; Lancet, D.; Kedem, O.; Pilpel, Y. *Origins of Life and Evolution of the Biosphere* **1998**, *28*, 501–514.

(8) Segré, D.; Ben-Eli, D.; Lancet, D. *Proceedings of the National Academy of Sciences of the United States of America* **2000**, *97*, 4112–7.

(9) Markovitch, O.; Lancet, D. *Artificial life* **2012**, *18*, 243–66.

(10) Orgel, L. E. *PLoS biology* **2008**, *6*, e18.

(11) Wu, M.; Higgs, P. G. *Journal of molecular evolution* **2009**, *69*, 541–54.

(12) Tkachenko, A. V.; Maslov, S. **2014**, 21.

(13) Nowak, M. A.; Ohtsuki, H. *Proceedings of the National Academy of Sciences* **2008**, *105*, 14924–14927.

(14) Ohtsuki, H.; Nowak, M. A. *Proceedings. Biological sciences / The Royal Society* **2009**, *276*, 3783–90.

(15) Chen, I. A.; Nowak, M. A. *Accounts of chemical research* **2012**, *45*.

(16) Derr, J.; Manapat, M. L.; Rajamani, S.; Leu, K.; Xulvi-Brunet, R.; Joseph, I.; Nowak, M. A.; Chen, I. A. *Nucleic acids research* **2012**, *40*, 4711–22.

(17) von Kiedrowski, G. *Angewandte Chemie International Edition in English* **1986**, *25*, 932–935.

(18) Lincoln, T. A.; Joyce, G. F. *Science* **2009**, *323*, 1229–1232.

(19) Vaidya, N.; Manapat, M. L.; Chen, I. A.; Xulvi-Brunet, R.; Hayden, E. J.; Lehman, N. *Nature* **2012**, *491*, 72–7.

(20) Shock, E. L. Stability of peptides in high-temperature aqueous solutions. 1992.

(21) Martin, R. B. *Biopolymers* **1998**, *45*, 351–353.

(22) PAECHT-HOROWITZ, M.; BERGER, J.; KATCHALSKY, A. *Nature* **1970**, *228*, 636–639.

(23) Leman, L.; Orgel, L. E.; Ghadiri, M. R. *Science (New York, N.Y.)* **2004**, *306*, 283–6.

(24) Orgel, L. E. *Critical reviews in biochemistry and molecular biology* **2004**, *39*, 99–123.

(25) Rao, M.; Odom, D. G.; Oró, J. *Journal of Molecular Evolution* **1980**, *15*, 317–331.

(26) Lambert, J.-F. *Origins of life and evolution of the biosphere : the journal of the International Society for the Study of the Origin of Life* **2008**, *38*, 211–42.

(27) Bernal, J. D. *Proceedings of the Physical Society. Section B* **1949**, *62*, 597–618.

(28) Ferris, J. P.; Hill, A. R.; Liu, R.; Orgel, L. E. *Nature* **1996**, *381*, 59–61.

(29) Nelson, K. E.; Robertson, M. P.; Levy, M.; Miller, S. L. *Origins of Life and Evolution of the Biosphere* **2001**, *31*, 221–229.

(30) Kanavarioti, A.; Monnard, P.-A.; Deamer, D. W. *Astrobiology* **2001**, *1*, 271–281.

(31) Bada, J. L. *Earth and Planetary Science Letters* **2004**, *226*, 1–15.

(32) Szostak, J. W.; Ellington, A. D. *The RNA World*; Cold Spring Harbor Laboratory Press, 1993; pp 511–533.

(33) Rode, B. M.; Son, H. L.; Suwannachot, Y.; Bujdak, J. **1997**, 273–286.

(34) Rode, B. M. *Peptides* **1999**, *20*, 773–786.

(35) Flory, P. J. *Principles of polymer chemistry*; Ithaca, NY : Cornell Univ., 1953; p 688.

(36) Stribling, R.; Miller, S. L. *Origins of Life and Evolution of the Biosphere* **1987**, *17*, 261–273.

(37) Huber, C.; Wächtershäuser, G. *Science (New York, N.Y.)* **1998**, *281*, 670–672.

(38) Aubrey, a. D.; Cleaves, H. J.; Bada, J. L. *Origins of Life and Evolution of Biospheres* **2009**, *39*, 91–108.

(39) Lazcano, A.; Miller, S. L. *Cell* **1996**, *85*, 793–798.

(40) Ding, P. Z.; Kawamura, K.; Ferris, J. P. *Origins of Life and Evolution of the Biosphere* **1996**, *26*, 151–171.

(41) Ferris, J. P. *Biological Bulletin* **1999**, *196*, 311.

(42) Chan, H. S.; Dill, K. A. *The Journal of Chemical Physics* **1991**, *95*, 3775.

(43) Sievers, A.; Beringer, M.; Rodnina, M. V.; Wolfenden, R. *Proceedings of the National Academy of Sciences* **2004**, *101*, 7897–7901.

(44) Lau, K. F.; Dill, K. A. *Macromolecules* **1989**, *22*, 3986–3997.

(45) Miller, D. W.; Dill, K. A. *Protein science : a publication of the Protein Society* **1995**, *4*, 1860–73.

(46) Yue, K.; Dill, K. A. *Proc Natl Acad Sci U S A* **1995**, *92*, 146–150.

(47) AGARWALA, R.; BATZOGLOU, S.; DANČÍK, V.; DECATUR, S. E.; HANNENHALLI, S.; FARACH, M.; MUTHUKRISHNAN, S.; SKIENA, S. *Journal of Computational Biology* **1997**, *4*, 275–296.

(48) Yue, K.; Dill, K. a. *Proceedings of the National Academy of Sciences of the United States of America* **1992**, *89*, 4163–4167.

(49) Xiong, H.; Buckwalter, B. L.; Shieh, H. M.; Hecht, M. H. *Proceedings of the National Academy of Sciences of the United States of America* **1995**, *92*, 6349–53.

(50) Fisher, M. a.; McKinley, K. L.; Bradley, L. H.; Viola, S. R.; Hecht, M. H. *PLoS ONE* **2011**, *6*, e15364.

(51) Giugliarelli, G.; Micheletti, C.; Banavar, J. R.; Maritan, A. *Journal of Chemical Physics* **2000**, *113*, 5072–5077.

(52) Ghosh, K.; Dill, K. A. *Proceedings of the National Academy of Sciences of the United States of America* **2009**, *106*, 10649–54.

(53) Stachelhaus, T.; Mootz, H. D.; Bergendahl, V.; Marahiel, M. A. *Journal of Biological Chemistry* **1998**, *273*, 22773–22781.

(54) Cherny, I.; Korolev, M.; Koehler, A. N.; Hecht, M. H. *ACS synthetic biology* **2012**, *1*, 130–8.

(55) Murphy, G. S.; Greisman, J. B.; Hecht, M. H. *Journal of molecular biology* **2015**, *428*, 399–411.

(56) Dill, K. A.; Bromberg, S.; Yue, K.; Chan, H. S.; Ftebig, K. M.; Yee, D. P.; Thomas, P. D. *Protein Science* **2008**, *4*, 561–602.

(57) Bryant, R. A. R.; Hansen, D. E. *Journal of the American Chemical Society* **1996**, *118*, 5498–5499.

(58) Smith, R. M.; Hansen, D. E. *Journal of the American Chemical Society* **1998**, *120*, 8910–8913.

(59) Ghosh, K.; Dill, K. A. *Biophysical Journal* **2010**, *99*, 3996–4002.

(60) Dill, K. A.; Ghosh, K.; Schmit, J. D. *Proceedings of the National Academy of Sciences of the United States of America* **2011**, *108*, 17876–82.