Prebiotically plausible mechanisms increase compositional diversity of nucleic acid sequences

Supplementary Information

Julien Derr, Michael L. Manapat, Sudha Rajamani, Kevin Leu, Ramon Xulvi-Brunet, Isaac Joseph, Martin A. Nowak, Irene A. Chen

Contents

1. Compositional diversity and structure formation

We develop a simple model to illustrate why compositional diversity is correlated with formation of secondary structure. Let us compute the probability of forming a hairpin in a sequence of length $L = 50$, for annealing length $k = 4$. Given H_4 of the sequence, we calculate the probability of finding one subsequence of length k in the 3' half of the sequence which is complementary to any subsequence of the 5' half.

Let Ω_1 and Ω_2 be the number of different k-length subsequences in the 3' and 5' halves of the sequence, respectively. The probability of forming a hairpin is:

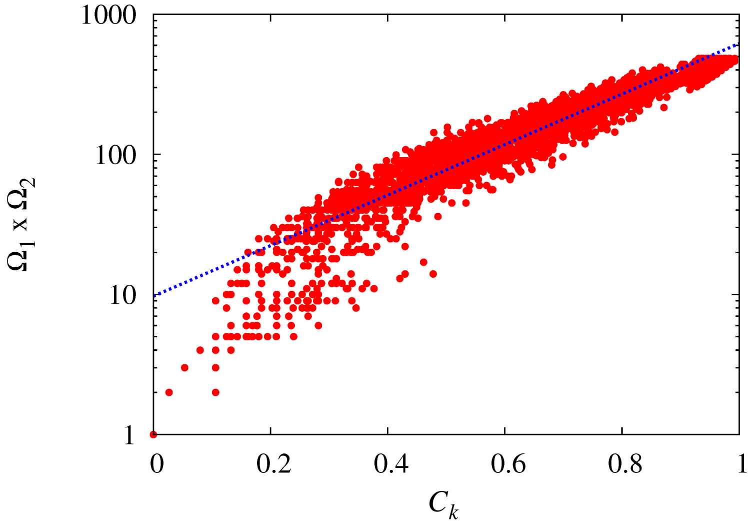
$$p_{hairpin} = 1 - \bar{p}_{hairpin} \tag{1}$$

in which

$$\bar{p}_{hairpin} = \prod_{i=1}^{\Omega_1} p_i \tag{2}$$

where i explores all the different subsequences of the 5' half, and p_i is the probability of not finding the complement of subsequence i on the 3' half. If $N = 4$ is the number of possible bases:

$$p_i = \prod_{j=1}^{\Omega_2} \left(1 - \frac{1}{N^k}\right) \tag{3}$$



Supplementary Figure 1: Exponential relationship between $\Omega_1 \times \Omega_2$ and C_k, with $k = 4$. The dashed line shows a linear fit: $\log(\Omega_1 \times \Omega_2) = aC_4 + b$, where $a \approx 4.2$, $b \approx 2.3$.

Therefore,

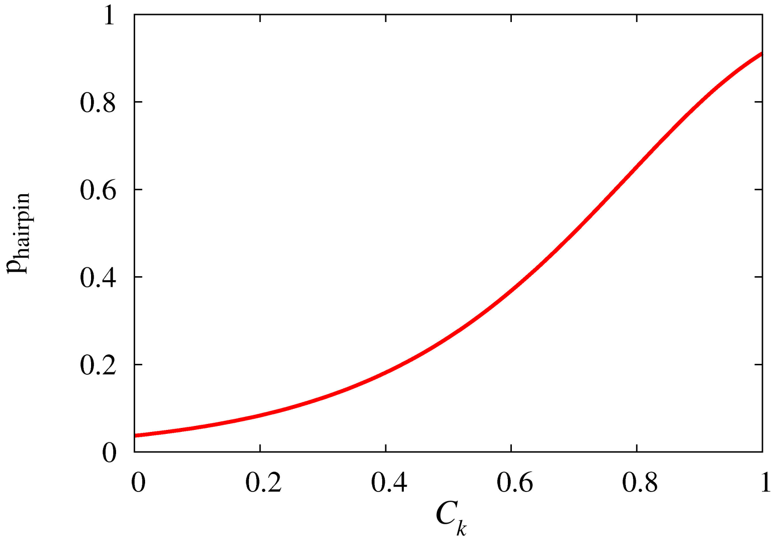
$$\bar{p}_{hairpin} = \left(1 - \frac{1}{N^k}\right)^{\Omega_1 \times \Omega_2} \tag{4}$$

Ω can be roughly linked to the normalized compositional diversity by the formula: $C_k = \log_2(\Omega)$. We verified this relationship numerically (Supplementary Figure 1), so $\log(\Omega_1 \times \Omega_2) = aC_k + b$. In terms of compositional diversity:

$$p_{hairpin} = 1 - \left(1 - \frac{1}{N^k}\right)^{e^{aC_k + b}} \tag{5}$$

Supplementary Figure 2 displays this equation graphically.

To the extent that more complicated secondary structures comprise multiple stem substructures, this reasoning would also apply. For example, the formation of a hammerhead is analogous to the formation of three hairpins with a particular set of length constraints (manifest in Ω). The precise calculation would vary depending on the values of Ω and whether the structure contains mismatched regions. Our simple hairpin model is meant to demonstrate the statistical reason why C_k is correlated with calculated minimum folding energy, but there is substantial variation from other factors (39%). The fact that many secondary structures are considered together in this correlation probably contributes to the additional variation.



Supplementary Figure 2: Probability $p_{hairpin}$ of forming a hairpin versus C_k, following equation 5.

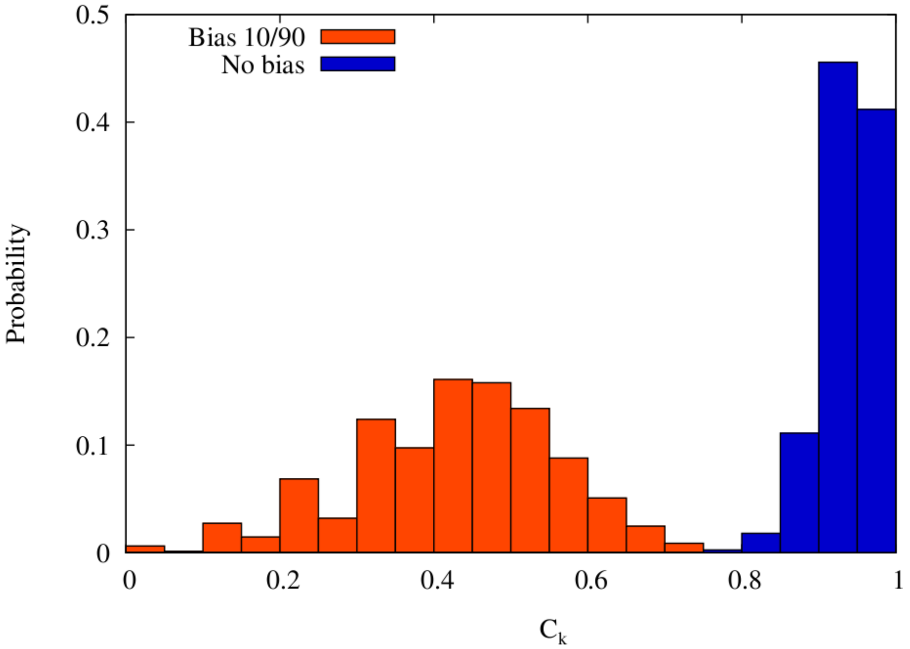
2. Biased composition restricts exploration of sequence space

We performed computer simulations to determine the influence that a typical bias has on the distribution of C_k. Each sequence was generated by randomly assigning a monomer value (0 or 1) to each position according to a Bernoulli process. In the unbiased case, each value has an equal probability of being chosen ($p = 1-p = 0.5$). For the biased case, the probability of incorporation of a 0 was $p = 0.9$ and the probability of incorporation of a 1 was $1-p = 0.1$. (This corresponds to a 9-fold bias in monomer abundance.) C_k was measured with $k = 4$. The results are shown in Supplementary Figure 3. The bias moves the distribution from $\langle C_k \rangle \approx 0.94$ to $\langle C_k \rangle \approx 0.43$. The drop is substantial, showing that achieving high compositional diversity in a biased environment is not trivial.

3. Parameters from experiments: c_0, r_{lig}, r_{con}

Our choice of parameter ranges was determined by searching the experimental literature and by our own work. The concentration of monomers (c_0) is an important parameter since it relates rate constants to the rate of reaction in the simulations. Typical values of the concentration used in experiments are:

- 20 mM nucleotides (template-directed polymerization from monomers [79])
- 100 mM nucleotides (template-directed polymerization from monomers [77])
- 25-50 mM nucleotide equivalents (template-directed ligation of hexamers to form dodecamers [84])



Supplementary Figure 3: Probability of generating a sequence of length 50 with a given C_4 in the biased case (red) and the unbiased case (blue).

- 50 mM monomers, 50 mM template (template-directed polymerization from monomers [52])
- 66 mM nucleotide equivalents (template-directed ligation from trimers to form hexamers [85])
- few mM nucleotide equivalents (template-directed ligation from random hexamers or dodecamers [34])
- 15 mM nucleotides (non-templated polymerization of monomers [21])

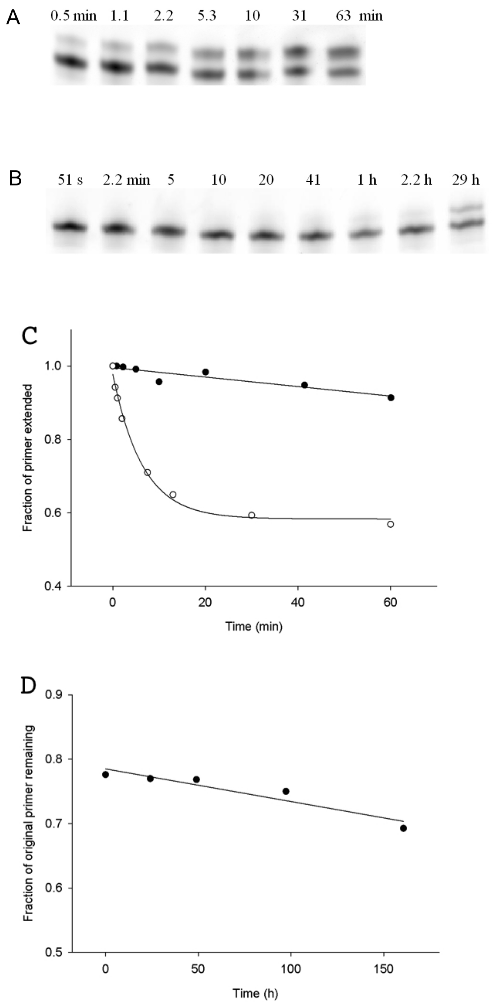
Therefore, we consider c_0 to be 0.01-0.1 M.

In general, $r_{\rm lig}$ depends on the activation chemistry and the rate of annealing [75] (Supplementary Table 1). The backbone conformation (A form vs. B form) appears to be less important in determining the rate of template-directed ligation, as shown by our comparison of an RNA template vs. DNA template (Supplementary Figure 4). Taking these different experimental systems into account, the relative strength of ligation to concatenation (r_{lig}) is on the order of 10^3 to 10^7.

The backbone strongly influences the rate of hydrolysis ($r_{\rm con}$), so for our modeling we focus on the range of parameters appropriate to RNA. Although $r_{\rm con}$ has not been

Reactants	$k_{\rm con}$	$k_{\rm lig}$	k_h	Reference
2-MeImpG ($\sim$ 100 mM nucleotide equivalents)	0.09 M^{-1} h^{-1}	430 M^{-2} h^{-1}		[52]
DNA trimers (carbodiimide activation; $\sim$ 4 mM nucleotide equivalents)	0.02 M^{-1} s^{-1}	1700 M^{-2} s^{-1}		[54]
Oligoribonucleotides (triphosphate)		1.3×10^5 M^{-2} h^{-1}	6×10^{-3} h^{-1}	[82]
ImpdG, DNA template ($\geq$10 mM nucleotide equivalents)	3±0.8 M^{-1} h^{-1}	$3 \pm 0.3 \times 10^8$ M^{-2} h^{-1}	$2 \pm 1 \times 10^{-5} h^{-1}$	This work, Supp. Fig. 4
ImpN, RNA template (10-40 mM nucleotide equivalents		3 x 10^7 to 10^9 M^{-2} h^{-1}		This work

1: Rate constants for concatenation, template-directed ligation, and hydrolysis.



Supplementary Figure 4: The constants k_{con}, k_{lig}, and k_h in a DNA analog. Polyacrylamide gel of the (a) templated reaction or (b) non-templated reaction over time. (c) Primer extension over time for templated (open circles) or non-templated (closed circles) reactions. (d) Primer degradation over time.

measured in a single system, we can infer it from the following: $r_{\mathrm{con}} = k_{\mathrm{con}} c_0 / k_{\mathrm{h}} = (k_{\mathrm{con}}/k_{\mathrm{lig}}) c_0{}^*(k_{\mathrm{lig}}/k_{\mathrm{h}})$, where $k_{\mathrm{con}}/k_{\mathrm{lig}}$ and $k_{\mathrm{lig}}/k_{\mathrm{h}}$ have been measured in RNA systems (Supplementary Table 1, first three lines). Therefore, we consider r_{con} to be on the order of 1 to 100. For our first approximation, we did not consider the effects of secondary structure. We also did not consider product inhibition, since this is less important for a promiscuous system [76, 78] and can be circumvented by spatial organization [80].

4. Stochastic Model

We begin with a collection of N_I monomers, N_0 of which are 0's and N_1 of which are 1's. The total number of monomers in the system, N_I, controls the volume of our model reactor. Thus, N_0/N_I and N_1/N_I are proportional to the initial concentrations of 0 and 1, respectively.

At any given time, the system consists of a variable number n of molecules (monomers and polymers), which we will denote by $P_1, \ldots, P_n$. At time 0, $P_1, \ldots, P_{N_0}$ are all equal to the monomer 0, and $P_{N_0+1}, \ldots, P_{N_I}$ are all equal to the monomer 1.

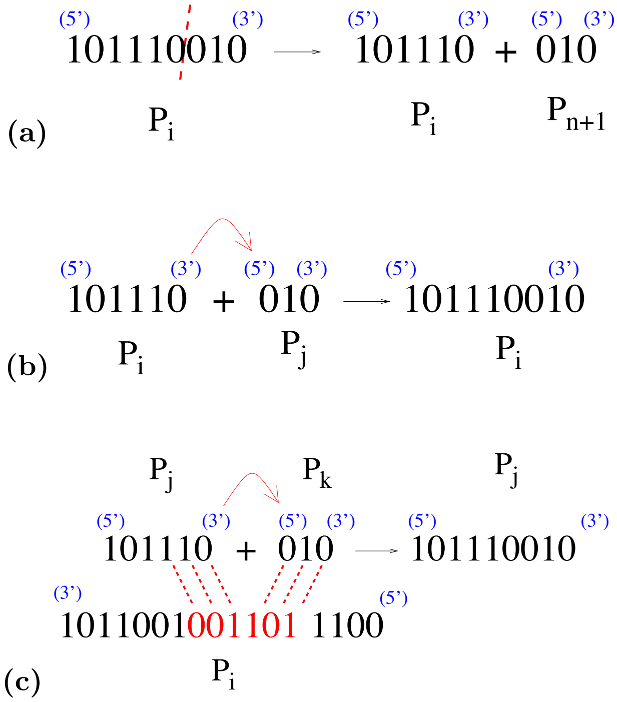
When the 5' monomer of a molecule is a 0, concatenation occurs with rate a. When the 5' site is a 1, concatenation occurs with rate b. Hydrolysis occurs with rate h (per bond, i.e., longer polymers have a greater chance of being hydrolyzed at some point along the chain). Template-directed ligation occurs with rate c.

Simulations of the system were based on the Gillespie algorithm. In each iteration, an exponential waiting time with parameter λ is generated, where λ is the sum of the rates of all the possible reactions. A particular reaction is then chosen to occur at random based on its rate relative to the other possible reactions. Suppose there are n polymers in the system at a given time. For a given polymer P_i, three reactions are possible (Supplementary Figure 5):

Hydrolysis of one of its bonds, each bond having an equal probability h of being hydrolyzed. If hydrolysis occurs, two fragments result. The first fragment remains labelled P_i, while the second fragment becomes the new polymer P_{n+1}.

Concatenation with another polymer, with rate a/N_I if the 5' site of P_j is a 0 and rate b/N_I if the 5' site of P_j is a 1. If concatenation occurs between P_i and P_j, the new polymer is labelled P_i, P_k is replaced by P_{k+1} for $k = j, \ldots, n-1$, and P_n is removed.

Template-directed ligation of two polymers, with rate $c/N_I^2 \times n_p$, where n_p is the number of potential ligation sites on the template, which depends on the complementarity between the template and the other available molecules in the pool. The extent of complementary required in these simulations is three consecutive bases at the 5' fragment, and three consecutive bases at the 3' fragment. If ligation happens using the template P_r and the two fragments P_s and P_t, P_s is extended to $P_s + P_t$, P_m is replaced with P_{m+1} for $m = t, \ldots, n-1$, and P_n is removed. The



Supplementary Figure 5: Schematic representation of the basic reactions of the model. **(a)** Hydrolysis. **(b)** Concatenation. **(c)** Template-directed ligation.

polymer P_i can participate in the ligation reaction either as a template or as one of the fragments that are joined.

To ensure that the reaction rates depend only on monomer concentration and not on the absolute number of monomers in the system, the rate constants in the model are normalized by N_I. The normalization is linear in N_I for concatenation but quadratic for template-directed ligation. This is described in the following section (Mapping simulation parameters with experimentally determined rate constants).

After each reaction, we measure various observables X of interest (e.g., the complexity). If we denote the exponential waiting time before the reaction by δt, this measurement makes a contribution of weight δt to the long-term average of X. To check when equilibrium has been reached, we measure the observables at successive, exponentially distributed time steps (times 0 and 1, then 1 and 2, then 2 and 4, etc). Steady-state is determined to have been reached when $(X_{t+1} - X_t)/X_{t+1}$ is less than a fixed tolerance (typically 0.1%).

5. Mapping simulation parameters with experimentally determined rate constants

We describe the link between the measured reaction rates (k_{lig}, k_{con}, k_{h}) and the computational parameters (a, b, c, and h).

Concatenation: We have the chemical reaction

$$A + B \xrightarrow{k_{\mathrm{con}}} AB, \quad \text{with} \quad \frac{d[AB]}{dt} = k_{\mathrm{con}}[A][B].$$

Let n be the number of polymers in the system, V the volume of the reactor, V_0 the average volume per monomer ($V_0 = 1/c_0$, where c_0 is the initial concentration of monomers), and N_I the total number of monomers. We can compare the individual rates in both the real and simulated systems:

$$k_{\mathrm{con}} \times \frac{n}{V} = \frac{B}{N_I} n = \frac{B}{V/V_0} n.$$

where B represents either a or b. The first expression corresponds to the real system and the second and third to the simulation. Therefore,

$$B = \frac{k_{\mathrm{con}}}{V_0}.$$

Template-directed ligation: The same reasoning applies for the equation

$$A + B + T \xrightarrow{k_{\mathrm{lig}}} AB + T, \quad \text{with} \quad \frac{d[AB]}{dt} = k_{\mathrm{lig}}[A][B][T].$$

We conclude that

$$c = \frac{k_{\mathrm{lig}}}{V_0^2}.$$

Hydrolysis: This is a first-order reaction, so it is not affected by c_0 (i.e., $h = k_{\mathrm{h}}$).

6. Deterministic Model

The deterministic model is the natural analogue of the stochastic one. As before, let a be the concatenation rate

$$i + 0j \rightarrow i0j$$

and b the concatenation rate

$$i + 1j \rightarrow i1j,$$

where i and j are any sequences (possibly null). Let h be the rate at which any given bond in a polymer is hydrolyzed.

We first formulate the system without ligation, so concatenation and hydrolysis are the only permissible processes. Let x_i denote the abundance of sequence i. We will now determine $\dot{x}_i$, the time derivative of the abundance.

Let P_i denote collection of all sequences that have i as a prefix and S_i the collection of all sequences that have i as a suffix. Then we recover i from hydrolysis of longer sequences at rate

$$h\left(\sum_{j\in P_i} x_j + \sum_{k\in S_k} x_k\right).$$

Let l be the length of i. There are $l-1$ bonds in sequence i, so i is lost due to hydrolysis at rate

$$h(l-1)x_i.$$

The total contribution to $\dot{x}_i$ due to hydrolysis is thus

$$h\left(\sum_{j\in P_i} x_j + \sum_{k\in S_k} x_k\right) - h(l-1)x_i.$$

Let $R_{i,0}$ denote the collection of all suffixes of i with leading bit 0, and for $m \in R_{i,0}$, let $L_i(m)$ be the prefix of i such that the concatenation of $L_i(m)$ and m (in that order) is i. Similarly, let $R_{i,1}$ denote the collection of all suffixes of i with leading bit 1, and for $n \in R_{i,1}$, let $L_i(n)$ be the prefix of i such that the concatenation of $L_i(n)$ and n (in that order) is i. Then i is formed by the concatenation of shorter sequences at rate

$$a\sum_{m\in R_{i,0}} x_{L_i(m)}x_m + b\sum_{n\in R_{i,1}} x_{L_i(n)}x_n.$$

Now let Z denote the collection of all sequence with leading bit 0, O the collection of all sequence with leading bit 1, and A the collection of all sequences. If the leading bit of i is 0, let $d = a$. Otherwise, let $d = b$. Then i is lost due to concatenation at rate

$$ax_i\sum_{p\in Z} x_p + bx_i\sum_{q\in O} x_q + dx_i\sum_{r\in A} x_r.$$

The total contribution to $\dot{x}_i$ due to concatenation is thus

$$a\sum_{m\in R_{i,0}} x_{L_i(m)}x_m + b\sum_{n\in R_{i,1}} x_{L_i(n)}x_n - ax_i\sum_{p\in Z} x_p - bx_i\sum_{q\in O} x_q - dx_i\sum_{r\in A} x_r.$$

When only hydrolysis and concatenation are possible, we therefore have

$$\begin{aligned}\dot{x}_i =& h\left(\sum_{j\in P_i} x_j + \sum_{k\in S_k} x_k\right) - h(l-1)x_i\\ &+a\sum_{m\in R_{i,0}} x_{L_i(m)}x_m + b\sum_{n\in R_{i,1}} x_{L_i(n)}x_n\\ &-ax_i\sum_{p\in Z} x_p - bx_i\sum_{q\in O} x_q - dx_i\sum_{r\in A} x_r.\end{aligned}$$

For the simulation to be computationally tractable, we must impose an arbitrary limit on sequence length, i.e., we do not allow there to be sequences longer than N. We can make the system above reflect this by (1) omitting sequences longer than N from all the collections and (2) omitting sequences longer than $n - l$ from the collections Z, O, and A (so those collections become dependent on i).

Now we add template-directed ligation to the system. Let c be the ligation rate. For each sequence i, let F_i denote the collection of triples (u, v, w) such that sequence u (the template) catalyzes the ligation of v and w to form i. (While the annealing length is not explicitly included in the formulation of the system, it determines which triples can be in F_i and is thus implicitly one of the parameters.) Then i is formed by ligation at rate

$$c \sum_{(u,v,w)\in F_i} x_u x_v x_w.$$

Let R_i denote the collection of all triples of the form (y, i, z) or (y, z, i), i.e., a triple in which i is one of the ligation reactants (but not the template). Then i is lost due to ligation at rate

$$c \sum_{(y,z,\alpha)\in R_i} x_y x_z x_\alpha,$$

where at least one of z and α is equal to i. When i acts as a template, its abundance does not change and thus there is no need to include its role as a template in the formulation of $\dot{x}_i$. The total contribution ligation makes to $\dot{x}_i$ is hence

$$c \sum_{(u,v,w)\in F_i} x_u x_v x_w - c \sum_{(y,z,\alpha)\in R_i} x_y x_z x_\alpha.$$

As before, these terms will be adjusted in practice to reflect the length limitation.

With hydrolysis, concatenation, and ligation, the full system is as follows:

$$\begin{array}{rl}
\text{time derivative of } i\text{'s abundance} & \left\{ \dot{x}_i = \right. \\
\text{formation by hydrolysis} & \left\{ h \left(\sum_{j\in P_i} x_j + \sum_{k\in S_k} x_k \right) \right. \\
\text{loss due to hydrolysis} & \left\{ -h(l-1)x_i \right. \\
\text{formation by concatenation} & \left\{ +a \sum_{m\in R_{i,0}} x_{L_i(m)} x_m + b \sum_{n\in R_{i,1}} x_{L_i(n)} x_n \right. \\
\text{loss due to concatenation} & \left\{ -a x_i \sum_{p\in Z} x_p - b x_i \sum_{q\in O} x_q - d x_i \sum_{r\in A} x_r \right. \\
\text{formation by ligation} & \left\{ +c \sum_{(u,v,w)\in F_i} x_u x_v x_w \right. \\
\text{loss due to ligation} & \left\{ -c \sum_{(y,z,\alpha)\in R_i} x_y x_z x_\alpha . \right.
\end{array} \tag{6}$$

We simulate this system until equilibrium has been reached and then compute the desired functions of the equilibrium distribution (average C_k, diversity, and length).

7. Size distribution

7.1. Without template-directed ligation, assuming $a = b$

We want to compute $p(l)$, the probability that a polymer randomly selected from the reactor has length l. In this simple case, we just have two parameters: B is the effective concatenation rate (see below) and h is the hydrolysis rate.

There are two ways of consuming a polymer of size l: concatenation (with rate $p(l)B$) or hydrolysis (with rate $(l-1)hp(l)$). There are also two ways of creating a polymer of size l: concatenation of smaller fragments (with rate $1/2\sum_{i=1}^{l-1} Bp(i)p(l-i)$) or hydrolysis of bigger fragments (with rate $2\sum_{i=l+1}^{\infty} hp(i)$). At steady-state the detailed balance equation

$$Bp(l) + (l-1)hp(l) = 1/2\sum_{i=1}^{l-1} Bp(i)p(l-i) + 2\sum_{i=l+1}^{\infty} hp(i). \tag{7}$$

should hold. If we rewrite the same detailed balance equation for polymers of length $l+1$ and then subtract the two equations, we obtain

$$\begin{aligned}[p(l+1) - p(l)](lh + B) + hp(l) = {} & \frac{B}{2}\sum_{i=1}^{l-1} p(i)[p(l+1-i) - p(l-i)] \\ & + \frac{B}{2}p(l)p(1) - 2hp(l+1).\end{aligned} \tag{8}$$

Based on our simulations, we guess an exponential solution of the form $p(l) = \alpha e^{\beta l}$, and we obtain

$$\alpha = \frac{2h}{B}, \tag{9}$$

$$\beta = -\log\left(1 + \frac{2h}{B}\right). \tag{10}$$

We can verify that this solution is properly normalized, i.e., that $S = 1$, where S is the sum of all probabilities:

$$S = \sum_{i=1}^{\infty} p(i) = \sum_{i=0}^{\infty} \alpha e^{\beta l} - \alpha = \alpha\frac{1}{1 - \frac{1}{1+\alpha}} - \alpha = 1. \tag{11}$$

The effective rate B reflects the average concentration of polymers in the pool. B is related to the absolute concatenation rate $B_0 = bh$ by the equation

$$B = B_0\frac{\langle n\rangle}{n_0}, \tag{12}$$

where $\langle n\rangle$ is the average number of polymers at steady-state and n_0 is the initial number of monomers. The average length of a polymer at steady-state is thus $\langle l\rangle = n_0/\langle n\rangle$. We can write $\langle l \rangle$ as a function of an infinite sum depending on α and β and express B accordingly:

$$B = B_0 \frac{1}{\alpha \sum_{i=1}^{\infty} i e^{\beta i}} = \frac{B_0 (1 - e^{\beta})^2}{\alpha e^{\beta}}. \tag{13}$$

We then rewrite equations (9) and (10) to obtain:

$$\begin{aligned} \alpha &= \frac{1}{X} - 1, \\ \beta &= \log(X). \end{aligned}$$

where X is the solution of $X^2 - X(2 + 2h/B_0) + 1 = 0$, with $X < 1$.

7.2. Without template-directed ligation, assuming $a \neq b$

The bias $a \neq b$ does not change the exponential shape of the distribution, but it does change the effective concatenation rate: $2aB$ is the concatenation rate for zeros, and $2bB = 2(1-a)B$ is the concatenation rate for ones. We can write the dynamics of the average population of zeros and ones (n_0 and n_1), assuming that hydrolysis gives as many reactive zeros as reactive ones (the term $\xi(n_0, n_1)$). The validity of this assumption is checked by the consistency of this analysis with the simulations (see below). The dynamics are:

$$\dot{n}_0 = -\frac{1}{2} 2aBn_0 + \xi(n_0, n_1), \tag{14}$$

$$\dot{n}_1 = -\frac{1}{2} 2(1-a)Bn_1 + \xi(n_0, n_1), \tag{15}$$

which give the steady state equilibrium

$$\frac{n_0}{n_1} = \frac{1-a}{a}. \tag{16}$$

We can relate this to the previous section by computing the effective concatenation rate

$$B_{\text{eff}} = 2B \frac{n_0 a + n_1(1-a)}{n_0 + n_1}, \tag{17}$$

which is equivalent to

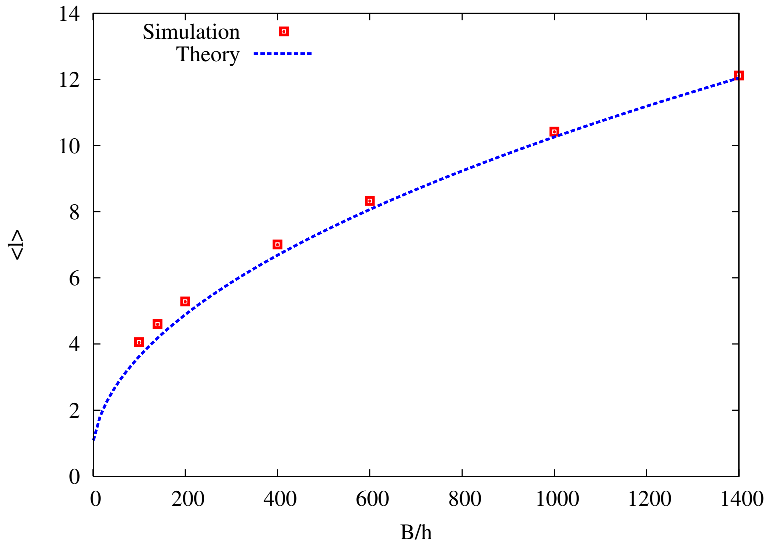
$$B_{\text{eff}} = 4Ba(1-a). \tag{18}$$

7.3. Average length without template-directed ligation

We can compute the average length, which we find to be $\frac{1}{1-X}$, or

$$\langle l \rangle = \frac{d}{\sqrt{1 + 2d} - 1}, \tag{19}$$

where $d = 4a(1-a)B_0/h$. This analytical calculation agrees very well with the results of the simulation (see Figure 6).



Supplementary Figure 6: The average length distribution for different concatenation rates (relative to hydrolysis) for a typically biased system ($a = 0.05$). The analytic calculations agree well with the simulations.

7.4. Size distribution with template-directed ligation

Rewriting the master equation (7) to include ligation gives

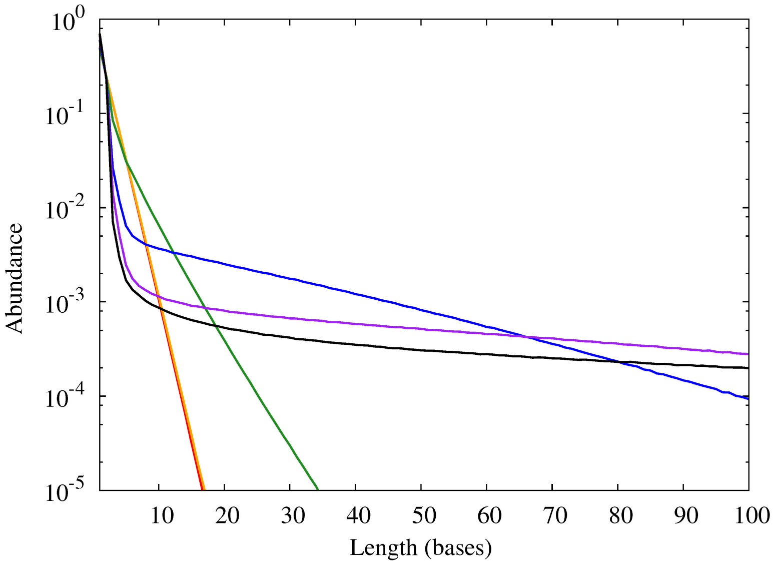
$$\begin{aligned} Bp(l) + (l-1)hp(l) = {} & \frac{1}{2}\sum_{i=1}^{l-1} Bp(i)p(l-i) + 2\sum_{i=l+1}^{\infty} hp(i) \\ & + \sum_{i \geq l_a; j \geq 2l_a} cf_{\mathrm{p}}p(i)p(j)(j-2l_a+1)p(l) \\ & + \frac{1}{2}\sum_{i=l_a}^{l-l_a}\sum_{j\geq 2l_a} cf_{\mathrm{p}}p(i)p(l-i)p(j)(j-2l_a+1), \end{aligned}$$

where l_a denotes the annealing length. The inclusion of the sum from l_a to $l - l_a$ in the sum from 1 to $l-1$ yields a constant term, so the solution is no longer exponential. This explains the skew of the resulting distribution; presumably the solution to this master equation would determine the new distribution.

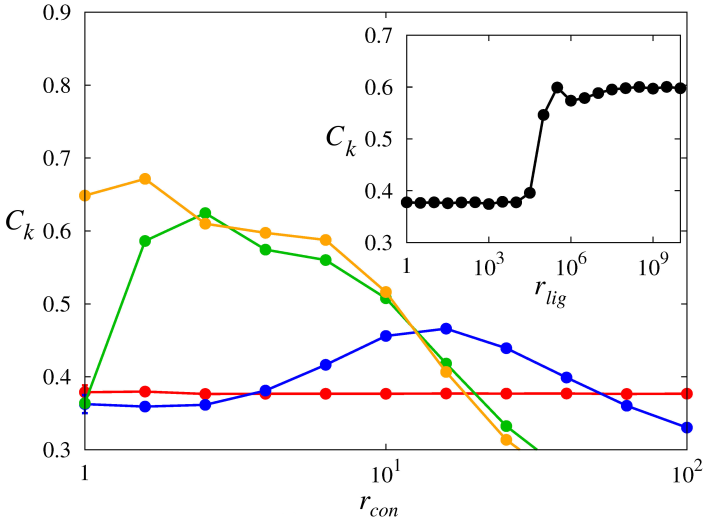
The length distribution in the presence of varying degrees of ligation is shown in Supplementary Figure 7).

8. Biased monomer abundance

When considering biased monomer abundance, template-directed ligation often increases C_k, but can also decrease C_k in some cases (Supplementary Figure 8). The C_k increase



Supplementary Figure 7: Template-directed ligation increases average length and complexity. Length distribution of binary sequences for different rates of template-directed ligation ($r_{lig} = 0$ (red), 10^2 (orange), 10^3 (green), 10^4 (blue), 10^5 (purple), 10^6 (black); $r_{con} = 10$ in all cases).



Supplementary Figure 8: Average C_3 versus concatenation ratio r_{con} of sequences of length 15 when monomer abundance is biased (initial ratio = 1:9; k=3), for r_{lig} = 0 (red), 2 x 10^4 (blue), 10^6 (green), or 2 x 10^8 (orange). The inset shows C_3 as a function of template-directed ligation (r_{con} = 4).

is likely due to factors described in the main text. The decrease occurs in a limited regime but indicates the presence of some competing effects.

One possible explanation for the favoring of repetitive sequences at very high ligation rates is that they may be more robust to hydrolysis. A low C_k sequence (mostly 1s) would be in presence of a complementary equivalent (mostly 0s). In this extreme parameter range, ligation dominates, and when hydrolysis is low ($k_{\mathrm{con}}c_0/k_{\mathrm{h}} \gg 1$), hydrolysis of one of these two polymers will be immediately repaired by template-directed ligation on the other. There are more ways for this to occur on a low C_k template, so this mechanism might therefore favor low C_k sequences.

The extent of this effect is apparently limited. As $k_{\mathrm{con}}c_0/k_{\mathrm{h}} \to \infty$, average C_k increases towards the same limit as for $k_{\mathrm{lig}} = 0$, as the system essentially comprises one polymer. While these parameters may not be biochemically relevant, this additional phenomenon illustrates the complicated effects that occur within even a relatively simple model.

9. Concatenation increases compositional diversity: a mass-action effect

Here we present a simple model that gives us an analytical understanding of the effect that concatenation rates have on C_k of sequences of a given length. Our model consists of two types of monomers, **0** and **1**, and two types of dimers, **00** and **11**. (We later expand this to include the heterodimers.) Monomers can be concatenated to form dimers, and dimers can be hydrolyzed into their constituent monomers:

$$\begin{aligned} \mathbf{0}+\mathbf{0} &\to \mathbf{00}, & \qquad \mathbf{00} &\to 2\cdot\mathbf{0}, \\ \mathbf{1}+\mathbf{1} &\to \mathbf{11}, & \qquad \mathbf{11} &\to 2\cdot\mathbf{1}. \end{aligned}$$

Concatenation of two $\mathbf{0}$'s to form the dimer $\mathbf{00}$ occurs at rate a. Concatenation of two $\mathbf{1}$'s to form the dimer $\mathbf{11}$ occurs at rate b. Both dimers are hydrolyzed at rate h. We assume that dimers are homogeneous in their constituents, but simulations suggest that relaxing this restriction does not affect the qualitative features we wish to establish with this model (see below).

We can formulate the chemistry described above with the following system of ordinary differential equations, where x_i is the abundance of sequence i:

$$\begin{aligned} \dot{x}_0 &= -2ax_0^2 + 2hx_{00} \\ \dot{x}_1 &= -2bx_1^2 + 2hx_{11} \\ \dot{x}_{00} &= ax_0^2 - hx_{00} \\ \dot{x}_{11} &= bx_1^2 - hx_{11} \\ 1 &= x_0 + 2x_{00} \\ 1 &= x_1 + 2x_{11} \end{aligned} \tag{20}$$

The last two equations guarantee that there are an equal number of $\mathbf{0}$'s and $\mathbf{1}$'s in the system.

At equilibrium, $\dot{x}_0 = 0$, so $ax_0^{*2} = hx_{00}^*$, where a $*$ denotes a quantity's value at equilibrium. Letting $a' = a/h$, we have $a'x_0^{*2} = x_{00}^*$. Since

$$x_0 + 2x_{00} = 1, \tag{21}$$

we obtain the relation

$$x_0^* + 2a'x_0^{*2} = 1. \tag{22}$$

Solving this yields

$$x_0^* = \frac{\sqrt{8a'+1}-1}{4a'}. \tag{23}$$

Similarly, we have

$$x_1^* = \frac{\sqrt{8b'+1}-1}{4b'}. \tag{24}$$

Thus,

$$\begin{aligned} \frac{x_0^*}{x_1^*} &= \frac{b'(\sqrt{8a'+1}-1)}{a'(\sqrt{8b'+1}-1)} \\ &= \left(\frac{b'}{a'}\right)\frac{(\sqrt{8a'+1}-1)(\sqrt{8b'+1}+1)}{8b'} \\ &= \left(\frac{b'}{8a'}\right)\left(\sqrt{\frac{64a'b'+8a'+8b'+1}{b'^2}} + \sqrt{\frac{8a'+1}{b'^2}} - \sqrt{\frac{8b'+1}{b'^2}} + \frac{1}{b'}\right). \end{aligned} \tag{25}$$

Now let $a', b' \to \infty$ while keeping the ratio a'/b' fixed. Then from the expression above, we see that

$$\frac{x_0^*}{x_1^*} \to \sqrt{\frac{b'}{a'}}. \tag{26}$$

When $b' > a'$, concatenation of **1**'s is faster than concatenation of **0**'s. Thus, there should be fewer free **1** monomers than **0** monomers, making x_0^*/x_1* large.

Since $x_{00} = a'x_0^2$ and $x_{11} = b'x_1^2$, when $a', b' \to \infty$ with a'/b' fixed, we have

$$\frac{x_{00}^*}{x_{11}^*} \to \frac{a'}{b'}\left(\frac{x_0^*}{x_1^*}\right)^2 = \frac{a'}{b'}\left(\sqrt{\frac{b'}{a'}}\right)^2 = 1. \tag{27}$$

We conclude that increasing the absolute concatenation rates drives the population of dimers towards an equal distribution of **00** and **11**. This is tantamount to saying that fast concatenation increases the population entropy of dimers. Indeed, the entropy of the population is maximized in the limit of infinitely fast concatenation.

Supplementary Figure 9**(a)** shows the percentage of monomers that have been incorporated into dimers at equilibrium as a function of the concatenation rate B/h when $a = 0.2$ and $b = 0.8$. This bias in reactivity favors the polymerization of 1's. The percentage of 0's in dimers is in light blue and the percentage of 1's in dimers is in dark blue. The ratio of these two quantities is in red. When this ratio is very small ($\ll 1$) or very large ($\gg 1$), the population distribution is skewed towards the dimer 00 or the dimer 11. When it is close to 1, there are a roughly equal number of 00's and 11's. As the figure shows, the ratio converges to 1 as $B/h \to \infty$.

Supplementary Figure 9**(b)** shows the analogous figure when the complete set of reactions is possible:

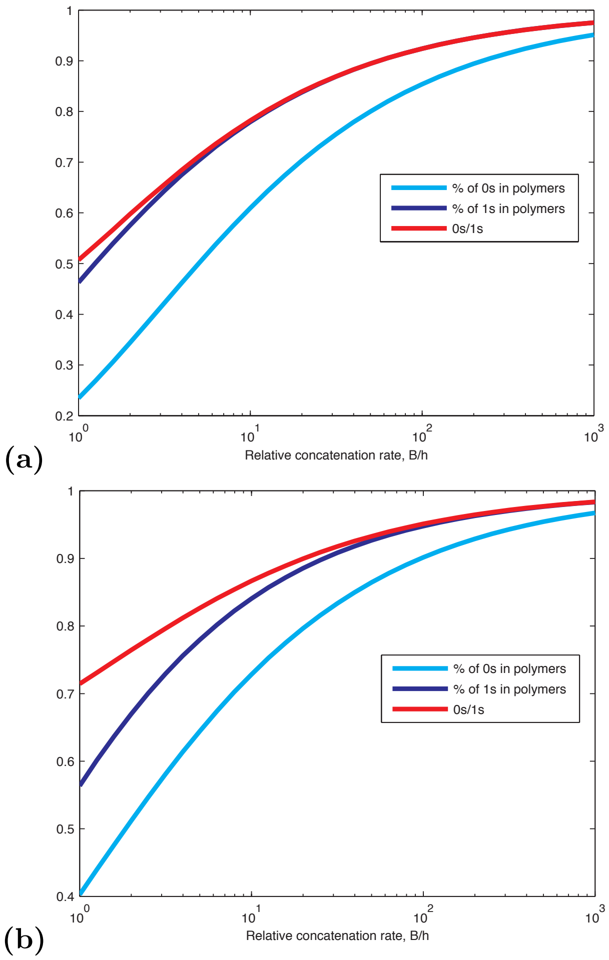
$$\begin{aligned}
&\mathbf{0} + \mathbf{0} \to \mathbf{00}, &\qquad& \mathbf{00} \to 2 \cdot \mathbf{0},\\
&\mathbf{1} + \mathbf{1} \to \mathbf{11}, && \mathbf{11} \to 2 \cdot \mathbf{1},\\
&\mathbf{1} + \mathbf{0} \to \mathbf{10}, && \mathbf{01} \to \mathbf{0} + \mathbf{1},\\
&\mathbf{0} + \mathbf{1} \to \mathbf{01}, && \mathbf{10} \to \mathbf{1} + \mathbf{0}.
\end{aligned}$$

In this case, the key quantity approaches 1 in the limit of high B/h as well.

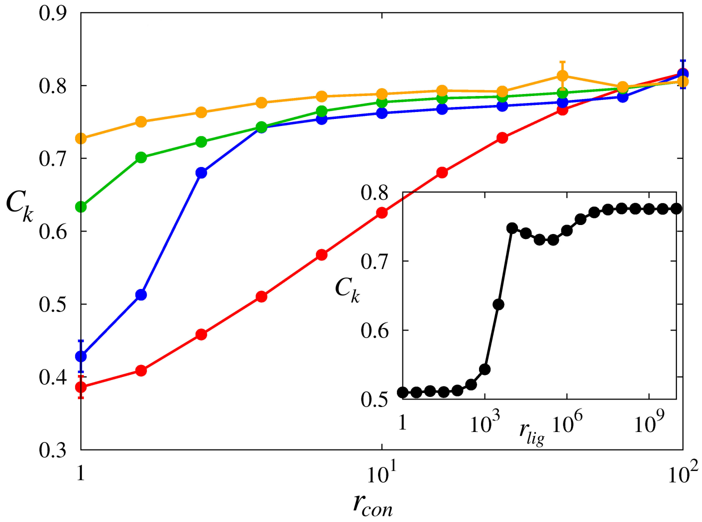
10. Compositional diversity and biased reactivity

When the reactivities of the two monomers differ, template-directed ligation increases compositional diversity (Supplementary Figure 10).

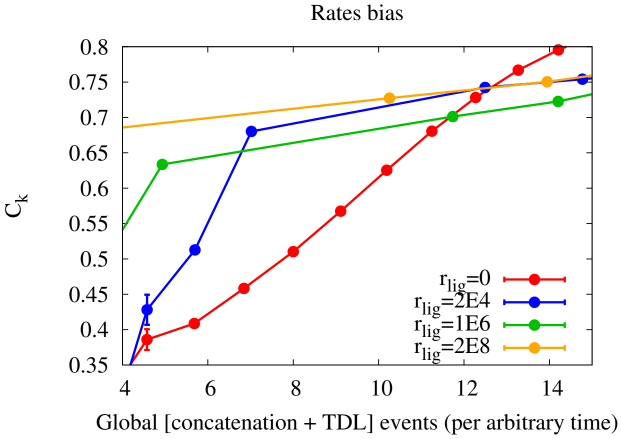
In order to check whether the effect of template directed ligation is just a mass action effect due to increased bond formation, we plotted (Supplementary Figure 11) C_k for different values of r_{lig}, as a function of the rate of bond forming events (template-directed ligation and concatenation). We observe that template-directed ligation still generally increases C_k, independently of its effect on bond formation rate, suggesting a more subtle cause (see main text).



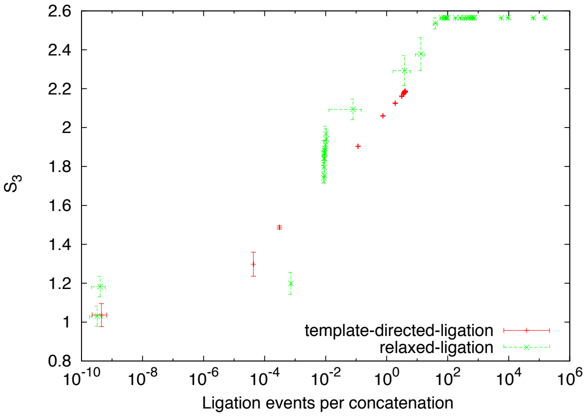
Supplementary Figure 9: The fraction of monomers that have been incorporated into dimers at equilibrium as a function of the concatenation rate B/h when $a = 0.2$ and $b = 0.8$. This bias in reactivity favors the polymerization of 1's. The percentage of 0's in dimers is in light blue and the percentage of 1's in dimers is in dark blue. The ratio of these two quantities is in red. Results are shown for the simplified system **(a)** and the full system **(b)**.



Supplementary Figure 10: C_3 versus r_{con} when monomer reactivity is biased (19-fold difference between k_{con}; r_{lig} = 0 (red), 2 x 10^4 (blue), 10^6 (green), or 2 x 10^8 (orange); k=3, length = 15). The inset shows C_3 as a function of template-directed ligation (r_{con} = 4).



Supplementary Figure 11: C_3 (length = 15) versus the total rate of bond-forming events, i.e., the sum of non- templated concatenation and template-directed ligation events.



Supplementary Figure 12: Average entropy (S_3) of 15-mers for systems with template-directed ligation (red) vs. relaxed-ligation (green), for simulation parameters giving a range of ratios for ligation events to concatenation events. Plotting the ratio of ligation events to concatenation events normalizes for the fact that relaxed-ligation produces many more ligation events overall compared to template-directed ligation.

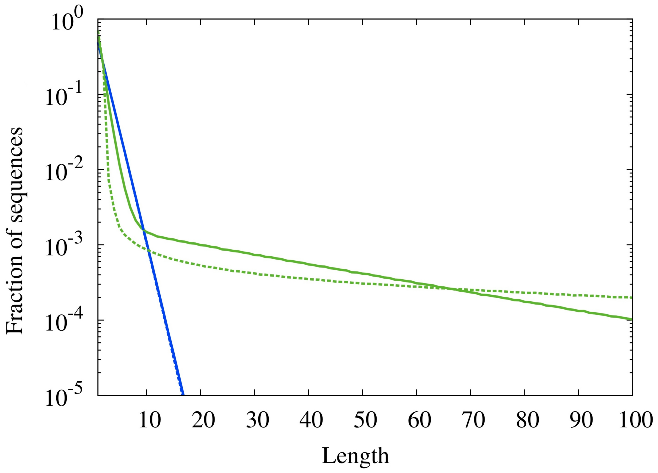
11. Template-directed ligation counters intrinsic bias

One role of template-directed ligation is to introduce a relatively unbiased mode of concatenation into the system (Supplementary Figure 12).

12. Stochastic simulation using a 4-letter alphabet

There are significant computational difficulties with repeating our simulations using a 4-letter system. For the deterministic simulations, the total number of different sequences that the program can keep track of is limited by computational resources, since every possible reaction among these sequences is computed. This imposes a practical limit on the maximum length of sequences permitted in the system. For a 2-letter system, the practical limit is a length of 12. For a 4-letter system, the practical limit would be a length of 6. We do not feel that a simulation up to this length is of interest, since 6 is also the realistic minimum length for a sequence to act as a template. Results from such a simulation could be easily misinterpreted due to length limitation effects. In our 2-letter simulations, we ran both stochastic and deterministic simulations to increase our confidence in the results. For the 4-letter simulation, we were only able to run stochastic simulations. Nevertheless, with this caveat, the results are described in this section.

We analyze the compositional diversity analysis of stochastic simulations with a 4-letter alphabet at $k = 1$ because the number of unique subsequences at $k > 1$ (i.e., 4^k) is larger than the number of subsequences analyzed per sequence for reasonable



Supplementary Figure 13: Length distribution for a 4-letter stochastic simulation (solid lines), with or without template-directed ligation ($r_{lig} = 0$ (blue) or 10^6 (green)). For comparison, the analogous results from a 2-letter simulation are shown in dotted lines (also shown in Figure 2a of the main text).

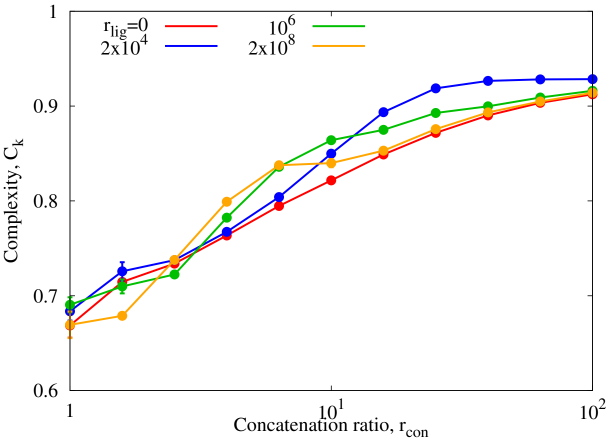
length. That is, we examine monomer composition of polymers of length 15 (longer polymers are rare, such that the number of simulations needed exceeds a reasonable computational capacity). The reactivity bias in the simulations was 19-fold (like Figure 2b of the main text), with two monomer types (e.g., the purines) reacting 10-fold faster than the other two monomer types (e.g., the pyrimidines). The length distribution is similar to the 2-letter case (Supplementary Figure 13). We observe a similar trend in compositional diversity, with template-directed ligation tending to increase C_1, although the effect is relatively small (Supplementary Figure 14). This is probably because the bias under study results in a relatively high compositional diversity even in the absence of template-directed ligation.

13. Why high C_k templates are favored

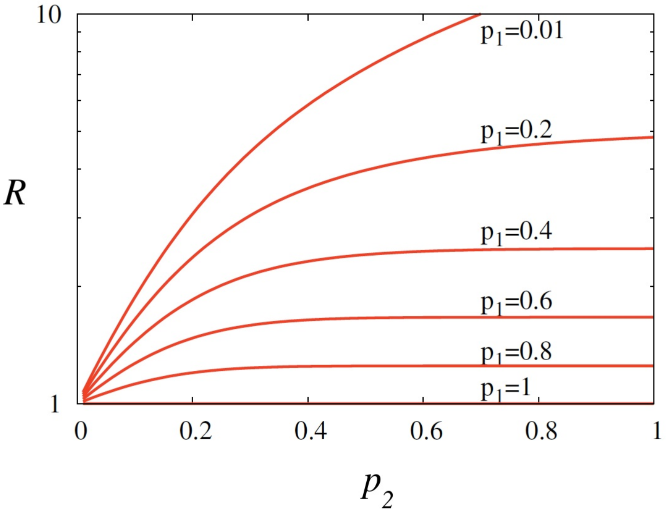
The ratio R is always greater than 1 (Supplementary Figure 15).

14. Cumulative frequency distribution of C_k for substrates and products of template-directed ligation (4 bases)

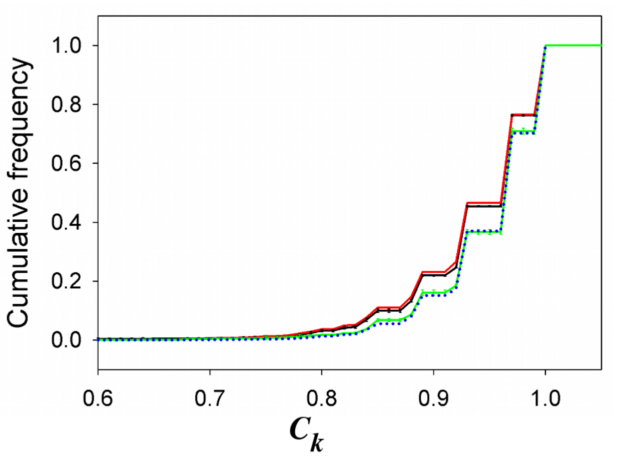
We found that the distribution of products of template-directed ligation was shifted toward higher C_k relative to the templates and octamer substrates. Supplementary Figure 16 shows the cumulative frequency distribution corresponding to the data described in Figure 3a of the main text. Note that the distribution of product sequences is shifted to the right. End-randomization did not affect this conclusion (Supplementary Figure 17).



Supplementary Figure 14: Average C_k ($k = 1$) for sequences of length 15 for a 4-letter stochastic simulation, over a range of concatenation ratio parameter values (r_{con}), with varying levels of template-directed ligation (r_{lig} = 0 (red), 2 x 10^4 (blue), 10^6 (green), 2 x 10^8 (yellow)).



Supplementary Figure 15: Relative templating ability (R) of high vs. low C_k templates, for different p_1 and p_2.

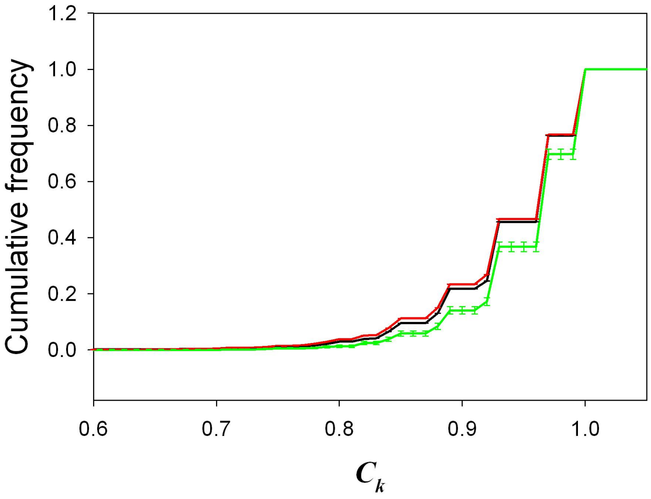


Supplementary Figure 16: Cumulative frequency distribution of average C_3 (length = 16) of template-directed ligation in a heterogeneous pool demonstrates that sequenced products (green) have increased C_3 relative to the templates (black = average C_3 of 16-mers contained in sequenced 40-mer templates) and octamers (red = C_3 of 16-mers from non-templated, random concatenation in silico from sequenced octamers), comparable to the distribution of a simulated completely random pool (dotted blue line). Error bars are standard deviations from replicate sequencing experiments.

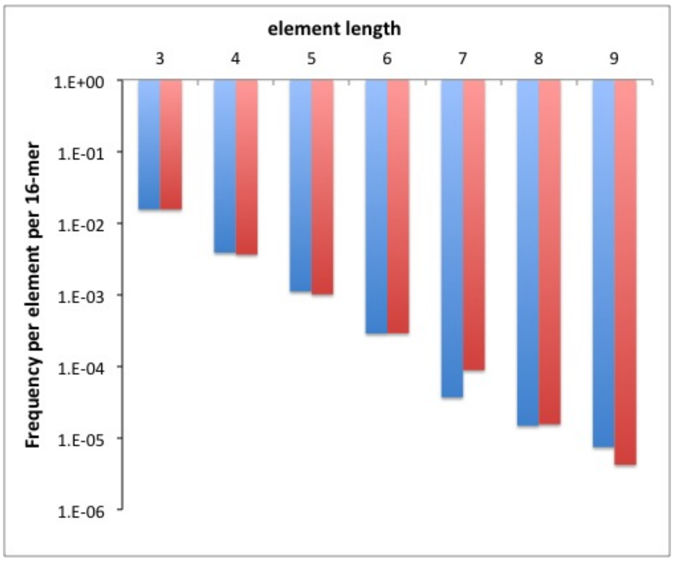
A systematic shift in the GC content also did not explain the change in C_3. The overall GC fraction of the sequence reads from templates, octamers, and ligation products was 0.50, 0.56, and 0.53, respectively.

15. Search for ribozyme elements in the products of experimental template-directed ligation

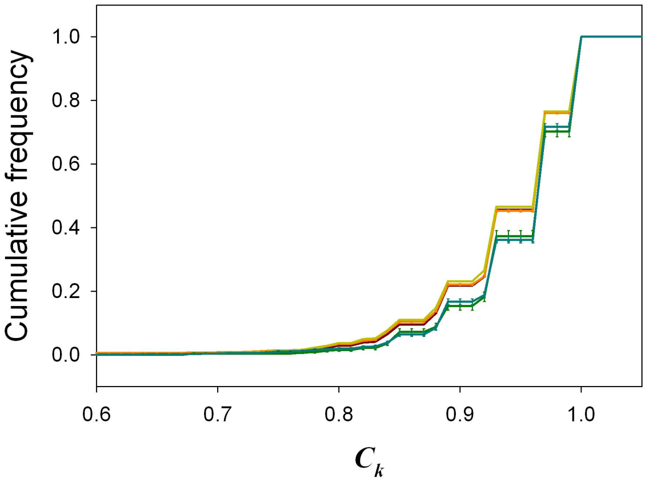
We were interested in whether ribozyme sequence elements are more highly represented in the products of experimental template-directed ligation compared to the products predicted from template-independent, random ligation of octamer substrates. We generated 10^6 16-mer sequences *in silico* by combining octamer sequences randomly chosen from a pool of 23353 octamers (octamer sequences were obtained experimentally). We refer to this set of predicted products of template-independent ligation as $\boldsymbol{P}$. We also sequenced 707 16-mer products of template-directed ligation; we refer to this set of experimental products as $\boldsymbol{E}$. Because the number of known ribozyme motifs is small, and likely to be much smaller than the set of all possible ribozyme motifs, we cannot directly search $\boldsymbol{P}$ and $\boldsymbol{E}$ for ribozyme motifs. Instead, we chose to use a ribozyme of particular importance for the RNA world theory, an RNA polymerase ribozyme (198 nt) [65]. We treated this ribozyme as a set of small sequence elements, and searched $\boldsymbol{P}$ and $\boldsymbol{E}$ for these elements, to determine whether the elements of this ribozyme are more or less frequent in $\boldsymbol{E}$ and $\boldsymbol{P}$. No obvious difference was seen (Supplementary Figure 18).



Supplementary Figure 17: Cumulative frequency distribution of C_3 for templates, octamers, and products, after randomization of first and last base of each sequence read. Black = template sequences; red = simulated 'products' from sequenced octamers; green = ligation product sequences. Error bars are from bootstrapping.



Supplementary Figure 18: Frequency of sequence elements from an RNA polymerase ribozyme [65] in the products of experimental template-directed ligation (red) or from simulated products of template-independent concatenation (blue). The frequency is the probability a particular sequence element was found in a given 16-mer, averaged over all elements.



Supplementary Figure 19: Cumulative frequency distribution of C_3 for templates, octamers, and products, showing inter-experiment differences and sampling error calculated by bootstrapping (error bars). Red and orange = template sequences from two experiments; yellow = simulated 'products' from sequenced octamers; green and blue = product sequences from two experiments.

16. Sampling error of sequencing (4 bases)

To calculate the sampling error from sequencing, 200 bootstrap samples were generated by randomly selecting n sequence reads (with replacement) from an experimental sample of n reads, and the compositional diversities were calculated for each sample. The sampling error is similar in magnitude to the error between experiments (Supplementary Figure 19).

17. Methods for experiments with binary templates

Template-directed chemical ligation with a limited subset of oligonucleotides (2 bases). We designed the following set of binary DNA templates and octamers such that a limited number (25, less than the theoretically possible set of 2^8) of octamers would be complementary to all 8mer subsequences within the entire set of templates, so that the availability of octamers would not be limiting. Ligation was performed with the following substrates. DNA oligonucleotides were obtained from Eurofins MWG Operon (Huntsville, AL) or Sigma-Aldrich (St. Louis, MO). For octamers, oligonucleotides were used without further purification. Template sequences were obtained in gel-purified form. Octamers were mixed in an equimolar mixture and phosphorylated by T4 polynucleotide kinase (New England Biolabs, Ipswich, MA) and [γ-^{32}P]-ATP at 37 degrees for 1 hour. 0.5 mM ATP was then added and the reaction was incubated overnight to ensure maximum phosphorylation. T4 PNK was inactivated by incubating at 65 degrees for 20 minutes.

```
Octamers:
5'-GGGGAGGA
5'-GGGAGGAG
5'-GGAGGAGA
5'-GAGGAGAA
5'-AGGAGAAA
5'-GGAGAAAA
5'-GAGAAAAG
5'-AGAAAAGA
5'-GAAAAGAA
5'-AAAAGAAG
5'-AAAGAAGA
5'-AAGAAGAG
5'-AGAAGAGG
5'-AAAAAAAA
5'-AGAGAGAG
5'-GAAGAGGA
5'-AAGAGGAG
5'-AGAGGAGA
5'-AGGAGAAG
5'-GGAGAAGA
5'-GAGAAGAG
5'-AGGAAGGA
5'-GGAAGGAA
5'-GAAGGAAG
5'-AAGGAAGG
```

```
32-mer templates:
T32:       5'-TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
Rep2_32T:  5'-CTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT
Rep4_32T:  5'-TCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCT
Rep8_32T:  5'-CCTCTTCTCCTCTTCTCCTCTTCTCCTCTTCT
Ran32T:    5'-CTCTTCTTTTCTCCTCTTCTTTTCTCCTCCCC
```

Oligonucleotides for measurement of product from a single template (2 bases). Random purine DNA octamers, `5'-RRRRRRRG` and `5'-RRRRRRRA`, where `R` denotes a purine, were mixed together in equimolar ratio and prepared as described above. Template oligonucleotide sequences were as follows:

```
40-mer templates:
T40:    5'-TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
2Tran:  5'-CTCTCTTCCCTTCTTTTTCCCCTTTTCTTTTCTCCTCCCC
```

```
Rep2T:  5’-CTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT
Rep6T:  5’-TCCTTCTCCTTCTCCTTCTCCTTCTCCTTCTCCTTCTCCT
Rep12T: 5’-TCTTCCTTTCCCTCTTCCTTTCCCTCTTCCTTTCCCTCTT
Rep20T: 5’-TCCTTCCCTTTTTTCCCCTCTCCTTCCCTTTTTTCCCCTC
3Tran:  5’-CCTCCTCTTCCTTTCCCCCTCCCTCTCTCTTTTCCTTCTC
```

Quantifying yield from single template reactions (2 bases). Molecular weight markers were obtained by radioactive phosphorylation of known oligonucleotides. Gels were exposed to a phosphorimaging screen (GE Healthcare, Piscataway, NJ) and the screen was scanned by a Typhoon TRIO Variable Mode Imager (Piscataway, NJ). Band intensity was quantified in ImageQuant. All band intensities were corrected for background measured from an unused portion of the gel. For quantitation, exposure times were limited to prevent over-saturation of signal intensity. Products were identified as bands with length 16 bases or more. Yield was calculated as the product band intensity divided by the sum of product band intensity and octamer band intensity.

18. Measurement of k_{con}, k_{lig}, and k_h in a DNA analog

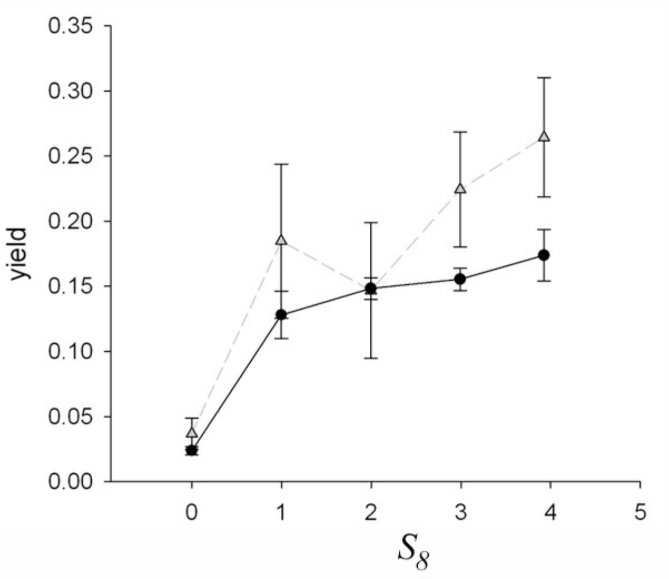
Experiments were performed according to Rajamani et al. [19]. A fluorescently labelled DNA primer terminated by a 3′-amino-2′,3′-dideoxynucleotide (5’-GGGATTAATACGACTCACTC-NH$_2$) was reacted with deoxyguanosine 5′-phosphorimidazolide (ImpdG) in the presence or absence of a template DNA oligonucleotide (5’-AGTGATCTCGAGTGAGTCGTATTAATCCC) to determine k_{con} and k_{lig}. The primer alone was incubated in the reaction buffer to determine k_{h}. The RNA templated reaction conditions were the same except for the primer sequence (5’-gggattaatacgactcactG-NH$_2$) and the template sequence (5’-agtgatctccagtgagtcgtattaatccc); lower case = RNA. Reactions were analyzed by denaturing polyacrylamide gel electrophoresis and quantified as previously described [19]. The initial rate of primer extension or degradation over time was fit to a straight line to obtain the apparent rate constant.

19. Template C_k and ligation yield

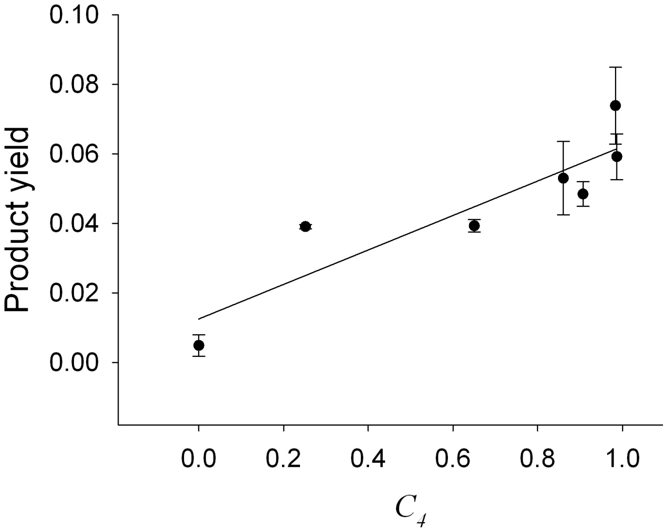
Templates with higher C_k gave greater yield from template-directed ligation. Supplementary Figure 20 shows quantitation corresponding to Figure 3b of the main text. Supplementary Figure 21 shows an analogous series with random binary sequences used as octamer substrates.

20. Diversity of the pool

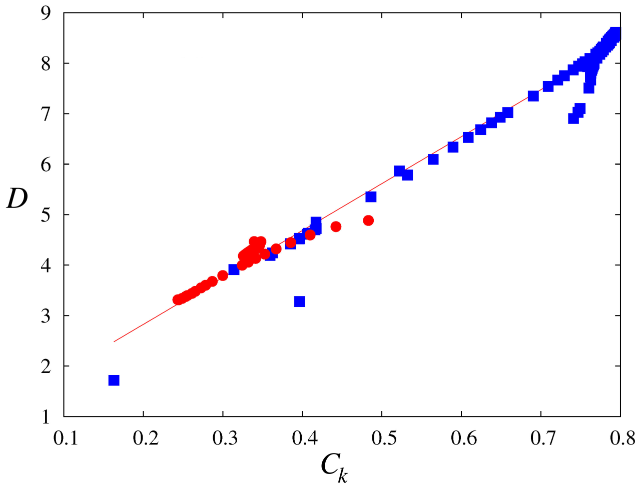
The diversity among sequences in the pool correlates well with the average complexity of sequences in the pool (Supplementary Figure 22).



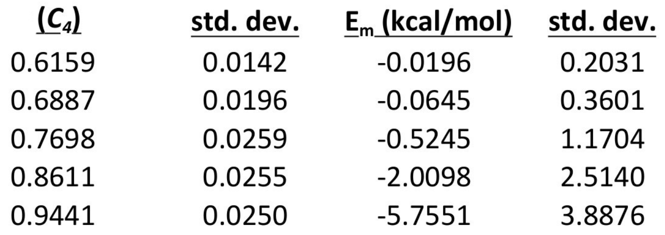
Supplementary Figure 20: Yield from template-directed ligation for a series of single templates of known C_k mixed with relatively high amounts of complementary fragments (32mer templates with subset of 25 possible octamers). Template concentration was 1 μM (black) or 2 μM (gray).



Supplementary Figure 21: Yield from template-directed ligation for a series of single templates of known C_k mixed with degenerate, random binary octamers (40mer templates).



Supplementary Figure 22: Diversity vs. average C_3 (length = 12) in deterministic simulations for r_{con} between 1 and 1000 and r_{lig} between 1 and 10^7 assuming bias in monomer abundance (red circles) or reactivity (blue squares). Solid line is a straight line of best fit to guide the eye (RMS deviation = 0.29).



(C_4)	std. dev.	E_m (kcal/mol)	std. dev.
0.6159	0.0142	-0.0196	0.2031
0.6887	0.0196	-0.0645	0.3601
0.7698	0.0259	-0.5245	1.1704
0.8611	0.0255	-2.0098	2.5140
0.9441	0.0250	-5.7551	3.8876

Supplementary Figure 23: Mean and standard deviations of C_4 vs. folding energy.

21. Folding energy and compositional diversity

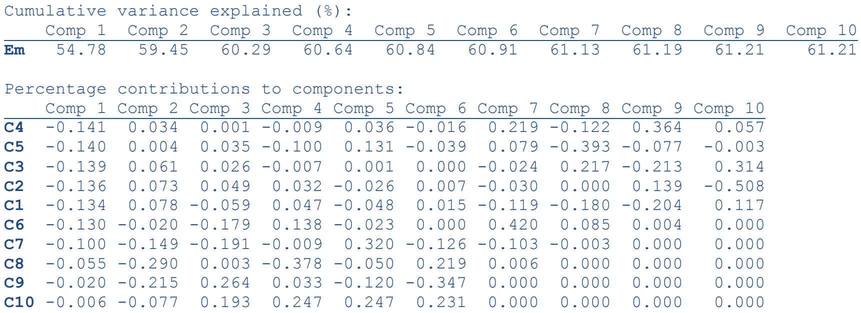
In Figure 1 of the main text, sequences were binned according to C_k, with bin average and standard deviation as given below (Supplementary Figure 23). The average folding energy and standard deviation are also given.

22. Ribozymes analyzed for main text Figure 1b

The compositional diversity of the following ribozymes was determined. These ribozymes were chosen from the Ellington Lab Aptamer Database if they had a length in the range of 40-60 bases (in order to compare the *in silico* folding of 50-mers).

```
————————————————————————————
Piganeau N, Thuillier V, Famulok M. J Mol Biol. 2001 Oct 5;312(5):1177-90.
————————————————————————————
sequences:
AAGGCTAGACTGCTAAGAGCGGAGTACCGTCATTGGTGTC ; c3=0.93982586976 ;c4=0.989623852973
AAGGCGAGACCGCTATGAGCGGAGTACCGTCATTGGTGTT ; c3=0.946069499894 ;c4=0.989623852973
AAGGCGAGACCGCTTTGAGCGGAGTACCGTCATTGGTGTT ; c3=0.936040478187 ;c4=0.989623852973
AAGGCAAGACCGCTATGAGCGGAGTACCGTCATTGGTGTT ; c3=0.959883913173 ;c4=0.989623852973
AAGGCAAGACCGCTATGAGCGGAGTACCGTCATCGGTGTT ; c3=0.949854891467 ;c4=0.989623852973
AAGGCCATACTTTGACTGATAGTCTTTGAGTACCGTTGTC ; c3=0.882109978081 ;c4=0.968871558918
CAGGCCATACTTGGACTGATTGTCCTTGAGTACCGTCGTC ; c3=0.902168021495 ;c4=0.979247705945
GGAGAGGCCAGAGGGAATACGATAGTCCCAGTACCGCGCTC ; c3=0.889609412919 ;c4=0.96991293488
GGAGAGGCCAGAGGGAATTCGATAGTCCCAGTACCGCGCTC ; c3=0.909014577997 ;c4=0.96991293488
GGAGAGGCCAGAGGGAATCGATAGTCCCAGTACCGCGCTC ; c3=0.905953413067 ;c4=0.968871558918
CAGGCTAGGCCCTATATATGATGCTGGGAGTACCGTCGTT ; c3=0.905953413067 ;c4=0.968871558918
AAGGCGAGATGGCCCTAAGCAAGACTAAGTACCGTCATCT ; c3=0.919767826347 ;c4=0.979247705945
GTAAGGCGAGACCATTCCATGGAGTACCGATTCAGCAGGC ; c3=0.889680761227 ;c4=0.968871558918
GTGACACATTGGCTTAATGTTGGGATAGTGCACCGGATAC ; c3=0.919767826347 ;c4=0.968871558918
GTGACACATTGGCTTAATGTTGGGATAGTGCACCGGATAC ; c3=0.919767826347 ;c4=0.968871558918
GTGACACATTGGCTTAATGTTGGGATAGTGCACCGGATAC ; c3=0.919767826347 ;c4=0.968871558918
GTGACACATTGGCTTAATGTTGGGATAGTGCACCGGATAC ; c3=0.919767826347 ;c4=0.968871558918
ACCGCTGTACCTTACCGGTATAGGACAGGCCATACTGAGG ; c3=0.888353608215 ;c4=0.979247705945
ACCGCTGTACCTTACCGGTATAGGGCAGGCCATACTGAGG ; c3=0.878324586509 ;c4=0.979247705945
GGAGAGACCACTTGAAAAAACAAGGTCGGTTATGTTTAGT ; c3=0.90973880464 ;c4=0.964955147062
GGAGAGGCCACTTGAAAAAACAAGGCCGGTTACGTTTAGT ; c3=0.899709782933 ;c4=0.954579000035
GGATACTGTTGTGTGCGAAGCATGATCCGCATACGTGGGC ; c3=0.902168021495 ;c4=0.968871558918
AAGAGCGCGACTGTAGAGGTCCTTAACAGTGTGCGGACTC ; c3=0.915982434774 ;c4=0.979247705945
CGGGAGAGTGCCCCAGGATTTTGGCAATCGTGTGAGGGTG ; c3=0.885895369654 ;c4=1.0
CAGGCTGCAGATGCTCACTTTAACGTTGAGATTGGCCGTC ; c3=0.919767826347 ;c4=0.989623852973
GGGCACAGCGCGTGTGTGTCATAACGCATATCTCTATGTC ; c3=0.855808304534 ;c4=0.944202853007
ATTATACTCATTTCCACTTAGTGGAGAGTCTGGTGAGATC ; c3=0.905953413067 ;c4=0.989623852973
ACCGGCTGGTGGAGTACAATTTGCCAGTGTAAGGCTAGAC ; c3=0.93982586976 ;c4=0.989623852973
CGTGAGTGAGGGCTAACATGTGTCTAGCTACAGTATTTAC ; c3=0.882109978081 ;c4=0.968871558918
ATTTGCGGTGTGACGGGGTCCTTCGGTCCCGGTATAGTGC ; c3=0.855808304534 ;c4=0.954579000035
GGACACAACCGTGACATTAAATCTAACTGGGATGTCGGCC ; c3=0.936040478187 ;c4=0.989623852973
AAGTATGTCTGTGTTCTTGAGAACATGATGACGCAGCTCT ; c3=0.904626260056 ;c4=0.989623852973
TATGAAGGACCAGGGCACGACGCTGTGATTACCCTTCGTC ; c3=0.949854891467 ;c4=1.0
GGAGGCACCGCCTCCTGGCAAGGATTCATATTGCTGGCTT ; c3=0.895924391361 ;c4=0.968871558918
ACTAGAGGGCGTGGACACGACGTGTGATATTCGCCGCTGT ; c3=0.926011456481 ;c4=0.989623852973
AGGCACGACGCGATGCTTCCGAGAGAAACCTGACGGTGCC ; c3=0.912197043201 ;c4=0.979247705945
GGGTTTAGCCCTGGCAAGCTTAGATATTGCTAGCTCTGTT ; c3=0.898382629922 ;c4=0.968871558918
GTGGTTGCATACCTACGTGCTTGTTAGGCTGCGTGGACAC ; c3=0.892138999788 ;c4=0.979247705945
TGGTGTATGAGCCAGGTATCCTCTGGGTGCCTCAAATCGC ; c3=0.895924391361 ;c4=0.968871558918
————————————————————————————
average c4=0.976770986 for 39 sequences
————————————————————————————
————————————————————————————
Kawazoe N, Teramoto N, Ichinari H, Imanishi Y, Ito Y. Biomacromolecules. 2001 Fall;2(3):681-6.
————————————————————————————
sequences:
UGAACGAGGGCGGAUGUAGAACAGGGGCUGGAAUGUUCGGGAUUUUCUG ; c3=0.856320541348 ;c4=0.960642913743
AUUCGUCUGUUGUGGCGGAGGAGGGUGAGUAGGUGUGGUUGAAGUGGAUCG ; c3=0.817756532033 ;c4=0.944960509656
UGGUUGUGACUUCAGGGAAAGGAUGAGCGGAGGACUCCUGAAUCUUAUGAUCCG ; c3=0.869258103766 ;c4=0.979259831248
CUCCUCUGUCGGCGGAGGUCAGGUUGUGCAGGGUCAACGAUGAGGAGCGAUU ; c3=0.857315065613 ;c4=0.949113258985
————————————————————————————
average c4=0.958494128 for 4 sequences
————————————————————————————
————————————————————————————
Mobley EM, Pan T. Nucleic Acids Res. 1999 Nov 1;27(21):4298-304.
————————————————————————————
sequences:
AAAACAAACUGAUCGAACGUCACGGUCCGCCACCCAGCUCUUCACUGCCCCCCCC ; c3=0.842162919035 ;c4=0.940592840387
AGCUCUAUCGCCUGACACAACGGUAUGACUGCCCCCGUGCCCACCCCCCCC ; c3=0.811085996467 ;c4=0.904313426265
AUAUCGAGACUCCCAAAUGUUUCUGGUGACGGGUCUCUUCAACUCAGUCCACCUCCUCUG ; c3=0.869425036475 ;c4=0.969922526215
GACAUACCACAGACACAUUGUAGUGGCUAGAGUGGCAAAUGACUUCAGCAUGCAGGUCCC ; c3=0.87338361458 ;c4=0.963907031458
ACACACCCUCUGGGUUGGAGCUCUCUAGCCACUGCGAACUCUUCACUCGCUUUUCGCUCC ; c3=0.826291951225 ;c4=0.93984505243
AGUCGCAUCCUGGACUUGGGCCGUCUUGGACGUGCGACCAGACCAUCGCUGACGUUGAUG ; c3=0.853208507121 ;c4=0.93984505243
————————————————————————————
average c4=0.943070988 for 6 sequences
————————————————————————————
————————————————————————————
Li J, Zheng W, Kwon AH, Lu Y. Nucleic Acids Res. 2000 Jan 15;28(2):481-8.
————————————————————————————
sequences:
TTTTGTCAGCGACTCGAAATAGTGTGTTGAAGCAGCTCTA ; c3=0.882109978081 ;c4=0.979247705945
```

```
TTAGTTCTACCAGCGGTTCGAAATAGTGAAATGTTCGTGA ; c3=0.872080956375 ;c4=0.923450558953
CAAAGATGCCAGCATGCTATTCTCCGAGCCGGTCGAAATA ; c3=0.929796848053 ;c4=0.989623852973
CAAAGATGCCTGCATGCTATTCTCCGAGCCGGTCGAAATA ; c3=0.926011456481 ;c4=0.989623852973
GTCTCCGAGCCGGTCGAAATAGTCAGGTGTTTCTATTCGG ; c3=0.905953413067 ;c4=1.0
CTTCTCCGAGCCGGTCGAAATAGTAGTTTTTAGTATATCT ; c3=0.868295564802 ;c4=0.954579000035
AGGTGTTGGCTGCTCTCGCGGTGGCGAGAGGTAGGGTGAT ; c3=0.852022912961 ;c4=0.954579000035
————————————————————————
average c4=0.97015771 for 7 sequences
————————————————————————
————————————————————————
Roth A, Breaker RR. Proc Natl Acad Sci U S A. 1998 May 26;95(11):6027-31.
————————————————————————
sequences:
CGGGTCGAGGTGGGGAAAACAGGCAAGGCTGTTCAGGATG ; c3=0.885895369654 ;c4=0.979247705945
AGGATTAAGCCGAATTCCAGCACACTGGCGGCCGCTTCAC ; c3=0.915982434774 ;c4=0.989623852973
————————————————————————
average c4=0.984435779 for 2 sequences
————————————————————————
————————————————————————
Tang J, Breaker RR. Proc Natl Acad Sci U S A. 2000 May 23;97(11):5784-9.
————————————————————————
sequences:
AUGCAAUGCAUUUGAGAACUGUAAGUUGUAUGAGGGCAUG ; c3=0.865837326241 ;c4=0.948119264863
AUGUGAUGCAUUUGAGAACUGCAAGUUGUAUGAGGGCAUG ; c3=0.862051934668 ;c4=0.968871558918
UUGCAAUGCCUUUGAGAACUGAAAGUUGUAUUAGGGAGUG ; c3=0.926011456481 ;c4=1.0
AUGCAUUGCGUUUGAGAACUGGAAGUUGAAUGAGGGCAUG ; c3=0.854481151523 ;c4=0.968871558918
GUGCAAUGCAUUUGAGAACUGUGAGUUGUAUUAGGUCAUG ; c3=0.899709782933 ;c4=0.979247705945
AUGUAAUGCAUUUGAGAACUCAAAGUUGUAUUAGGGCAUG ; c3=0.915982434774 ;c4=0.979247705945
AUGCAAUUCAUUUGAGAACCGUAAGUUGUAUCAGGGCAUG ; c3=0.929796848053 ;c4=1.0
AUGCGAGGCAUUUGAGAACUUCAAGUUGUAUGAGGGCAUG ; c3=0.892138999788 ;c4=0.95849541189
AUGCAUUGCACUUGAGAGCUGAAAGCUGGAUGAGGGCAGG ; c3=0.858266543095 ;c4=0.95849541189
GUGCAAUGCAUUUGAGAACUGGAAGUUGUAUUAGGGCAUA ; c3=0.926011456481 ;c4=0.979247705945
AUGCUAUGCAUUUGAGAACUGAGAGUUGUAUGGGGGCACG ; c3=0.868295564802 ;c4=0.948119264863
AUGCAGUGCGUUUGAGAACUGUAAGUUGUAUCAGGGCAGG ; c3=0.895924391361 ;c4=0.968871558918
GUGCAAUGCAUUUGAGAACUGAAAGUUGUAUUAGGGUAUG ; c3=0.90973880464 ;c4=0.979247705945
GGCGAUAGGUGAGUACACUGGGUCGGAGGGAUAGCUAGGU ; c3=0.878324586509 ;c4=0.95849541189
GGCGAUAAGUGAGCACACUGGGUCGGAGGGCUAGCUAGGU ; c3=0.899709782933 ;c4=0.979247705945
GGCGAUAGGUGAGUACGCUGGGUCGGAGGGAUAGCUAGGC ; c3=0.872080956375 ;c4=0.968871558918
GGCGAUAAGUGAGUACACUGGGUCGGAGGGAUAGCUAGGA ; c3=0.905953413067 ;c4=0.989623852973
GGCGAUAGGUGAAUACACUGGGUCGGAGGGACGGCUAGGC ; c3=0.892138999788 ;c4=0.989623852973
GGCGAUAAGUGAGUACACUGGGUCGGAGGGAUAGCUAGGU ; c3=0.90973880464 ;c4=0.989623852973
GGCGAUAAGUGGGUACACUGGGUCGGAGGGAUAGGGAGGU ; c3=0.848237521389 ;c4=0.927366970808
GGCGAUAGGUGAAUACACUGGGUCGGAGGGAUAGCUAACA ; c3=0.905953413067 ;c4=0.979247705945
GGCGAUAGGUGAGUACACUGGGUCGGAGGGAUAUCGAGGU ; c3=0.868295564802 ;c4=0.968871558918
GGCGAUAGGUGAGUACAGUCGGUCGGAGGGAUUGCUAGUC ; c3=0.872080956375 ;c4=0.968871558918
GGCGAUAAGUGAGAACACUGCGUCGGAGGGAUCGCUAGGU ; c3=0.93982586976 ;c4=1.0
GGCCAUAGGUGUUUACACUGGGUCGGAGGGAUAGCUAAGC ; c3=0.929796848053 ;c4=0.989623852973
GGCGAUAGGUGAGUACACUGGGUCGGAGGGGUAGGAAGAU ; c3=0.878324586509 ;c4=0.979247705945
————————————————————————
average c4=0.974059632 for 26 sequences
————————————————————————
————————————————————————
Jadhav VR, Yarus M. Biochemistry. 2002 Jan 22;41(3):723-9.
————————————————————————
sequences:
AUUCGUCGAGGAGCUCACCAGGACUUAAUAAGUGCCAGUGCGCCGCUUCC ; c3=0.888092355156 ;c4=0.969356369931
AUUCGUCGAGGAGCUCACCAGGGCUUAAUAAGUGCCAGUGCGCCGCUUCC ; c3=0.885276432406 ;c4=0.969356369931
UAUUUCGUCGAGGACCAUAGCAUGUCGUAAAACAAUGACAAGGCGCUUCC ; c3=0.907657961375 ;c4=0.977017277448
CUCCGUCGAGGAACGAUGCAUCGAACAUAGAUUAGACAUCGUCGCUUCCC ; c3=0.860078980687 ;c4=0.938712739862
AGAGAUCCGUCGAGGACAGUUGUAUACCAGAGUGAGCAGCUGCGCUUCC ; c3=0.902285986452 ;c4=0.992128582749
AGAGAUCCGUCGAGGACAGUUGUAUAACAGAGUGAGCAGCUGCGCUUCC ; c3=0.894625078935 ;c4=0.984257165497
ACUGGCAUAACUCUUUGGGCAUGUGCGUCAGACCACGUGUUACCGCCAGC ; c3=0.917934393781 ;c4=0.984678184965
AAUAAAGGCAAUGGACAUAUCCAUCCCAGGAAGCCCCUGCGCCUCCUUGC ; c3=0.87218407725 ;c4=0.992339092483
AAAGAAAAUUCAAAGACAGGGCGUGGAGGAAUAUCCUGGAACUCUUUGCC ; c3=0.897381528969 ;c4=0.969356369931
UCCCGAUUGCAAUGACCUGCUCAUGGGCUAAACCCAAUUUUAGCUCGCG ; c3=0.91283845564 ;c4=0.984257165497
UCAGUGAAAGGUACCUCUCAAAUGUGAUCGAGGCAUUGUUUAAUGCAGGC ; c3=0.900197451719 ;c4=0.977017277448
ACAGAUACUCAAACGAAUAGUCUUAGCAAUUGGAACUUUAUAUACUCCG ; c3=0.874533918053 ;c4=0.968514330995
GUACGGAUCAGAAAAUGAAGAAACAUCCUCCGAUGGGGUGCAUAAUCUGC ; c3=0.887105096563 ;c4=0.984678184965
UCAGCCCCAUUACAUCGAUAUGCAAAUCACUUGAGGGUCUUAAGUCGUG ; c3=0.938712739862 ;c4=1.0
AACUACUAAAUGCGUUUCCGUCGAGGAUAUUCAGAAUCGAAUACGCUUCC ; c3=0.897381528969 ;c4=0.969356369931
GAAAAAUAACCAUAUCUUCCAAGAAUGCAAUCAGGGCUCAUUACAUUUGG ; c3=0.894565606219 ;c4=0.984678184965
————————————————————————
average c4=0.977856479 for 16 sequences
————————————————————————
————————————————————————
Beaudry A, DeFoe J, Zinnen S, Burgin A, Beigelman L. Chem Biol. 2000 May;7(5):323-34.
————————————————————————
sequences:
GGUGUCAUCAUAAUGGCACCCUUCAAGGACAUCGUCCGGG ; c3=0.932255086615 ;c4=0.979247705945
GGAGUCAUCAUAAUGGCUCCCUUCAAGGACAUCGUCCGGG ; c3=0.912197043201 ;c4=0.979247705945
GGUGUCAUCAUAAUGGCACCCUUCAAGGACAUAGUCCGGG ; c3=0.932255086615 ;c4=0.979247705945
GGAGCCAUCAUAAUGGCUCCCUUCAAGGACAUCGUCCGGG ; c3=0.936040478187 ;c4=0.989623852973
GGUGUCAUCAUAAUGACACCCUUCAAGGACAUCGUCCGGG ; c3=0.912197043201 ;c4=0.968871558918
```



```
Cumulative variance explained (%):
     Comp 1  Comp 2  Comp 3  Comp 4  Comp 5  Comp 6  Comp 7  Comp 8  Comp 9  Comp 10
Em    54.78   59.45   60.29   60.64   60.84   60.91   61.13   61.19   61.21    61.21

Percentage contributions to components:
     Comp 1 Comp 2 Comp 3 Comp 4 Comp 5 Comp 6 Comp 7 Comp 8 Comp 9 Comp 10
C4   -0.141  0.034  0.001 -0.009  0.036 -0.016  0.219 -0.122  0.364   0.057
C5   -0.140  0.004  0.035 -0.100  0.131 -0.039  0.079 -0.393 -0.077  -0.003
C3   -0.139  0.061  0.026 -0.007  0.001  0.000 -0.024  0.217 -0.213   0.314
C2   -0.136  0.073  0.049  0.032 -0.026  0.007 -0.030  0.000  0.139  -0.508
C1   -0.134  0.078 -0.059  0.047 -0.048  0.015 -0.119 -0.180 -0.204   0.117
C6   -0.130 -0.020 -0.179  0.138 -0.023  0.000  0.420  0.085  0.004   0.000
C7   -0.100 -0.149 -0.191 -0.009  0.320 -0.126 -0.103 -0.003  0.000   0.000
C8   -0.055 -0.290  0.003 -0.378 -0.050  0.219  0.006  0.000  0.000   0.000
C9   -0.020 -0.215  0.264  0.033 -0.120 -0.347  0.000  0.000  0.000   0.000
C10  -0.006 -0.077  0.193  0.247  0.247  0.231  0.000  0.000  0.000   0.000
```

Supplementary Figure 24: Principal components regression of C_k and E_m.

```
GGAGUCAUCACAAUGGCUCCCUUCAAGGACAUCGUCCGGG ; c3=0.905953413067 ;c4=0.989623852973
GGAGACAUCAUAAUGGCUCCCUUCAAAGACAUCGUCCGGG ; c3=0.915982434774 ;c4=0.95849541189
GGAGUCAUCAUUGUGGCUCCCUUCAAGGACAUUGUCCGGG ; c3=0.892138999788 ;c4=0.95849541189
GGUGCCACCAUAAUGGCACCCUUCAAGGACAUCGUCCGGG ; c3=0.959883913173 ;c4=0.989623852973
GGAGUCAUCAUAAUGGCUCCCUUCAAGGGCAUCGUCCGGG ; c3=0.902168021495 ;c4=0.979247705945
GGAGUCAUCAUAAUGGCACCCUUCAAGGACAUCGUCCGGG ; c3=0.922226064908 ;c4=0.979247705945
GGUGUCAUCAUAAUGACACCAUUCAAGGACAUCGUCCGGG ; c3=0.895924391361 ;c4=0.968871558918
GGUGUCAUCAUAACGACACCCCUCAAGGACAUCGUCCGGG ; c3=0.902168021495 ;c4=0.968871558918
GGUGUCAUCUUAAUGGCACCCUUCAAGGACAUCGUCCGGG ; c3=0.949854891467 ;c4=0.989623852973
GGAGCCGUCAUAAUGGCUCCCUUCAAGGACAUCGUCCGGG ; c3=0.929796848053 ;c4=0.989623852973
GGAGCCACCAUAAUGGCUCCCUUCAAGGACAUUGUCCGGG ; c3=0.959883913173 ;c4=1.0
GGAGUCAUCAUAAUGGCUCCCAUCAAGGACAUCGUCCGGG ; c3=0.882109978081 ;c4=0.954579000035
GGAGUCAUCAUAGUGGCUCCCUUCAAGGACAUCGUCCGGG ; c3=0.902168021495 ;c4=0.979247705945
GGAGUCACCAUAAUGGCUCCCUUCAAGGACAUCGUCCGGG ; c3=0.949854891467 ;c4=1.0
GGUGUCACCAUAGUGGCACCCUUCAAGGACAUCGUCCGGG ; c3=0.93982586976 ;c4=0.989623852973
GGUGUCAACAUAAUGACACCCUUCAAGGACAUCGUCCGGG ; c3=0.926011456481 ;c4=0.968871558918
GGAGUCACCAUAAUGACUCCCUUCAAGGACAUCGUCCGGG ; c3=0.93982586976 ;c4=1.0
GGUGUCAUCAUAAUGGCACCCUUCAAGGACAUCGUCUGGG ; c3=0.922226064908 ;c4=0.979247705945
GGAGUCACCAUAAUGGCUCCCUUCAAGGACAUCGUCCGGG ; c3=0.949854891467 ;c4=1.0
GGUGUCAUCAUAAUGGCUCCCUUCAAGGACAUCGUCCGGG ; c3=0.922226064908 ;c4=0.979247705945
GGUGUCAUCGUAAUGGCACCCUUCAAGGACAUCGUCCGGG ; c3=0.93982586976 ;c4=0.968871558918
GGAGUCAUCAUAAUGACUCCCUUCAAGGACAUCGUCCGGG ; c3=0.902168021495 ;c4=0.979247705945
GGUGUCAUCAUAAUGGCACCCAUCAAGGACAUCGUCCGGG ; c3=0.902168021495 ;c4=0.954579000035
GGUGUCAUCAUAAUGGCACCCAUCAAGGACAUCGUCCGGG ; c3=0.902168021495 ;c4=0.954579000035
GGAGUCAUCAUAAUGGCUCCCCUCAAGGACAUCGUCCGGG ; c3=0.892138999788 ;c4=0.979247705945
GGUGUCAUCAUAAUGGCACCCUUCAAGGACAUCCUCCGGG ; c3=0.922226064908 ;c4=0.979247705945
——————————————————————
average c4=0.978533984 for 31 sequences
——————————————————————

——————————————————————
```

23. Compositional diversity and folding energy: principal components regression

Nonparametric principal component regression [81] was carried out on ranked data using the R [83] package ”pls” [86] (Supplementary Figure 24).

24. References

See main text for references 75-86 cited in the supplementary information.