# Product inhibition can accelerate evolution

**Beatrice Ruth**[a,1,2] **and Peter Dittrich**[a,b,1]

[a]Department of Mathematics and Computer Science, Friedrich Schiller University Jena, 07744 Jena, Germany; [b]Jena Centre for Bioinformatics, 07743 Jena, Germany

This manuscript was compiled on June 14, 2022

**Molecular replicators studied in-vitro exhibit product inhibition, typically caused by the hybridization of products into dimer complex that are not able to replicate. As a result, the replication rate and the selection pressure is reduced, potentially allowing the "survival of everyone". Here, we introduce a stochastic evolution model of replicating and hybridizing RNA strands to study the effect of product inhibition on evolution. We found that hybridization, though reducing the rate of replication, can increase the rate of evolution, measured as fitness gain within a period of time. The positive effect has been observed for a mutation error smaller than half of the error threshold. In this situation, frequency-dependent competition causes an increased diversity that spreads not only within a neutral network but also over various neutral networks through a dynamical modulation of the fitness landscape, resulting in a more effective search for better replicators. The underlying model is inspired by RNA virus replication and the RNA world hypothesis. Further investigations are needed to validate the actual effect of accelerated evolution through product inhibition in those systems.**

self-replication| chemical evolution | RNA | fitness landscape | product inhibition

Replication of polymers is central to the reproduction of organisms and viruses(1), a key element in major theories of the origin of life(2), and of interest to make synthetic life (3, 4). Enzyme-controlled(5) and enzyme-free(6, 7) (self-)replication has been instantiated *in-vitro* and used to study chemical evolution(8, 9) and to implement bio-molecular procedure (10), like the polymerase chain reaction (11).

In *in-vitro* experiments of molecular replication, it has been observed that accumulation of product tends to inhibit the replication process, leading to sub-exponential growth(6, 7). In fact, product inhibition causes a reduction of the replication rate and reduction of selection pressure, which can lead to the survival of everyone(12).

For example, von Kiedrowski(6) has observed that the concentration of self-replicating hexadeoxynucleotides does not grow exponentially but sub-exponentially. The reason is the formation of hybrids, which act as sinks for single stranded molecules and thus limiting their growth. The effect has been theoretically investigated for replicating RNA sequences by Biebricher et al. (13). They also showed that different single stranded molecules can coexist without cooperative hypercyclic coupling, proposed by Eigen and Schuster(14).

Experimentally observations by Rohde et al. (15) showed that hybridization leads to a broad mutant distribution of an RNA species replicated by Q$\beta$ replicase. The models introduced in the context of the mentioned studies explain the coexistence of different types of species using rate equations where all molecular types are predefined. Such models do not cover how new types are formed and how mutation influences the time evolution of the population. In contrast our model enables the formation of new types through the simulation of single molecules, here single RNA strands. To avoid the problem of defining all molecular types in advance, our model uses an improved exact stochastic simulation algorithm similar to the Gillespie algorithm(16) instead of a set of ordinary differential equations.

Although product inhibition supports the "survival of everyone", an additional directed selection pressure can lead to certain adaptations(17) and can cause complex patterns of species formation (18). For example, Ito et al. (19) showed that adaptive radiation through intraspecific competition together with weak directional selection of a quantitative trait can lead to rich macroevolutionary patterns involving recurrent adaptive radiations and extinctions. If directional selection is sufficiently weak, evolutionary branching can occur under product inhibition (20). However, because of its narrow scope of evolvability, product inhibition and the resulting parabolic replication has been seen to be of limited relevance for prebiotic evolution (21), and thus mechanisms circumventing product inhibition have been suggested, like compartments (22) or the formation of intramolecular secondary structures (23).

Yet, the quantitative benefit on the fitness gain caused by product inhibition has not been studied in detail (21).

For an explicit simulation of RNA sequences evolution Fontana and Schuster(24) introduced a fitness function that uses the secondary structure of an RNA sequence to compute its fitness. We follow this approach, because it provides a certain level of realism while the fitness being efficiently computable in polynomial $O(n^3)$ time (25). The RNA fitness landscape possesses neutral networks spanning whole sequence space(26). Typical evolution is characterized by quasispecies distributions expanding and drifting on those neutral networks, improving in fitness by contineous and discontineous transitions(27, 28).

RNA viruses profiting from drifting and expanding on neutral networks (29, 30) may also experience product inhibition during replication in their hosts. The subsequent question than would be if the observed effect of our simulations is transferable to those viruses, contributing to an explanation for their high genotypic variance and quick adaption to new environments (29, 30).

## Results

**Model.** In our model [*], a well-stirred population of replicating and hybridizing RNA molecules is simulated. Replication requires a substrate $S$ with copy number $N(S)$, initially set to determine the maximal possible number of RNA molecules. A single stranded RNA molecule with sequence $r$ and copy number $N(r)$ replicates with error by consuming substrate $S$ in volume $V$ at a rate $\alpha(r)\frac{N(r)N(S)}{V}$, with replication rate constant $\alpha(r)$. A single stranded RNA molecule decays at a stochastic rate $\phi N(r)$, here $\phi = 1$, releasing substrate $S$. A sequence $r$ represents that sequence as well as its complement, simplifying the replication model.

Two RNA molecules $r, r'$ can hybridize at a rate $\beta(r, r')\frac{N(r)N(r')}{V}$ forming a complex. Note that in

[*] The code of the complete model is at https://git.uni-jena.de/ne78xoy/hr-sim-newgillespie.

### Significance Statement

In this paper we present a novel evolutionary phenomenon, where product inhibition, though reducing the effective replication rate, can accelerate the rate of evolution. We show this phenomenon in a model of simulated single-stranded RNA (sRNA) sequence evolution extended by hybridization of sRNA, causing product inhibition. The evolutionary phenomenon could be relevant in (a) prebiotic evolution, where replicating polymers hypothetically emerged and where very likely subject to product inhibition, (b) biotic evolution, e.g., where RNA strands of viruses replicate within a biological cell, or (c) artificial molecular or chemical evolution, where product inhibition might be used to evolve molecules with desired properties more efficiently.

The authors declare no competing interests.

[1] B.R.(Beatrice Ruth) contributed equally to this work with P.D. (Peter Dittrich).

[2] Correspondence should be addressed to Beatrice Ruth. E-mail: beatrice.flickuni-jena.de

case both sequences $r$ and $r'$ are equal the rate results from $\beta(r,r')\frac{N(r)(N(r')-1)}{2V}$(16). Furthermore, we assume that the resulting dimer complex cannot replicate and dissolving of the complex back into two single strands is slow such that it can be omitted. Thus the formed complex can be ignored. To maintain a constant maximal possible number of single stranded RNA molecules the complex is replaced by $2S$.

In summary, our model consists of the following reactions for RNA sequences $r, r' \in \{A, C, G, U\}^l$, which are similar to Epstein's non-reproductive pairing model(31):

$$r + S \xrightarrow{\alpha(r)} r + r' \quad \text{with} \quad r' = \text{mutate}(r) \quad , \tag{1}$$

$$r + r' \xrightarrow{\beta(r,r')} 2S \quad , \tag{2}$$

$$r \xrightarrow{\phi} S \quad . \tag{3}$$

Here, we simulate RNA sequences of fixed length $l = 76$ bases. The function mutate($r$) returns a mutated copy of $r$ where each site is mutated with probability $p$. For the replication rate constant $\alpha(r)$ we take a standard model of RNA evolution, namely the scaled distance $d_{sec}(r, r_{target})$ of the secondary structure of $r$ to a fixed target secondary structure $r_{target}$; an approach that is said to provide a relatively realistic fitness landscape. (27):In our case the target secondary structure is the shape of a tRNA (see Materials and Methods).

$$\alpha(r) = k_\alpha f_{\text{scaled}}(r), \tag{4}$$

$$f_{\text{scaled}}(r) = \frac{0.01}{1.01 - f(r)}, \tag{5}$$

$$f(r) = 1 - \frac{d_{\text{sec}}(\text{fold}(r), \text{fold}(r_{\text{target}}))}{l}. \tag{6}$$

The scaling factor $k_\alpha$ is set to 100. The process of folding a sequence $r$ into a secondary structure and the secondary structure alignment $d_{sec}(.,.)$ are computed by the *fold* and *tree_edit_distance* functions of the *ViennaRNA* package(32), respectively.

The hybridization rate constant $\beta(r, r')$ is computed from the hybridizing sequences $r, r'$ as

$$\beta(r, r') = k_\beta h_{\text{scaled}}(r, r'), \tag{7}$$

$$h_{\text{scaled}}(r, r') = \frac{h(r, r')^5}{0.55^5 + h(r, r')^5}. \tag{8}$$

Here, the hybridization strength $k_\beta$ is varied over the values $\{0, 0.1, 0.3, 1, 3, 4, 10\}$, where 0 implies no hybridization and increasing values lead to a stronger hybridization influence. The hybridization coefficient $h(r, r')$ depends on the hybridizing single strands $r$ and $r'$. Roughly, the more similar they are the more likely they hybridize. Because we assume RNA sequences with a fixed length $l$, we can compute the hybridization coefficient $h(r, r')$ from the sum of Gibbs free energy contribution of each base pair considering the adjacent base pairs (see Methods for details). Note that this leads to a more realistic model than using the Hamming distance, because the Watson-Crick base pairs CG UT have different contributions as they allow a different amount of hydrogen bonds.

For simulation, we usually generate a random initial population of $N(r) = 50$ RNA sequences of length $l = 76$ with $N(S) = 50$ substrate, allowing a total population of $n = 100$ RNA sequences, and let it evolve by an improved exact stochastic simulation algorithm similar to the Gillespie algorithm (for details and a proof of correctness see Methods). The new algorithm is necessary, because the standard Gillespie algorithm cannot cope with the large number of possible hybridization reactions, which scale quadratically with $n$.

Time $t$ is measured in generations. Technically, $t$ is incremented by one fraction of the current population size in each replication event (see Methods).

**Hybridization can accelerate evolution.** We measure the rate of evolution by the gain of fitness $f_{\mathrm{scaled}}$ of the population's best sequence after a period of time $\Delta t$.

At a low generation number $\Delta t$ the comparison of populations with and without hybridization revealed that a higher fitness level is only achievable without hybridization (data not shown). As time progresses ($\Delta t = 10000$) it is observable that populations with hybridization can have a higher efficiency in gaining fitness than populations without hybridization (Fig. 1). This effect is maintained over many generations ($\Delta t = 100000$) leading to a rising gap in reached fitness between populations with and without hybridization, from $\approx 0.01$ to $\approx 0.13$ a.u. (Fig. 1).

**Hybridization inhibits evolution when mutation is close to the error threshold.** The error threshold $p_{err}$ is a limit of the mutation probability of a base pair above which mutation will destroy the sequence information over time(14, 33). In our model, without hybridization ($k_\beta = 0$) we have $p_{err} \approx 0.02$ per base, in line with Kupczok & Dittrich (28).

Simulations performed with a high mutation rate of $p = 0.02$ showed that if the mutation probability is close to the error threshold, hybridization has always a negative effect on the rate of evolution (Fig. 2). This is inline with the fact that hybridization reduces the overall growth rate of a sequence and causing a bifurcation from a stable to an unstable regime. We can also see that with increasing hybridization strength the effective error threshold decreases.

**Decreasing population density can accelerate evolution under hybridization.** As high mutation rates have shown to have a negative effect on the fitness development of populations with hybridization, moderate mutation rates reveal a more benefiting evolution for populations with hybridization compared to populations without hybridization (Fig. 1). For such a moderate mutation rate, there is a regime of hybridization strengths ($1 < k_\beta < 4$) where increasing the volume leads to an increased rate of evolution (Fig. 3). Increasing the volume is equivalent to decreasing the population density by keeping the number of sequences and substrate constant. Leading to a decrease of replication and hybridization rates while the first order decay rate stays unchanged.

**Improvement by hybridization is caused by a broader mutant spectrum.** We measure population diversity and structure by Hamming distance among pairs of sequences. With increasing hybridization strength $k_\beta = 0$ to h $k_\beta = 10$ the mean Hamming distance increase roughly linearly from 12 to 40 bases (Fig. 4). Note that for $k_\beta = 10$ the mean Hamming distance is larger than the mean distance between two random sequences ($38 = l/2$). Further note that at such a high hybridization strength there is no evolutionary progress anymore (Figs. 1-3).

So, an improved rate of evolution through hybridization coincides with a a moderately broader mutant spectrum. Furthermore, we can see an increased number of clusters in sequence space (Fig. 5(b)) than within a population without hybridization (Fig. 5(a)). With its larger mutant spectrum the population with hybridization can explore the sequence space more effectively.

While a population without hybridization usually occupies one neutral network during its exploratory phase (27), a population with hybridization might occupy stably more than one neutral network (Fig. 5).

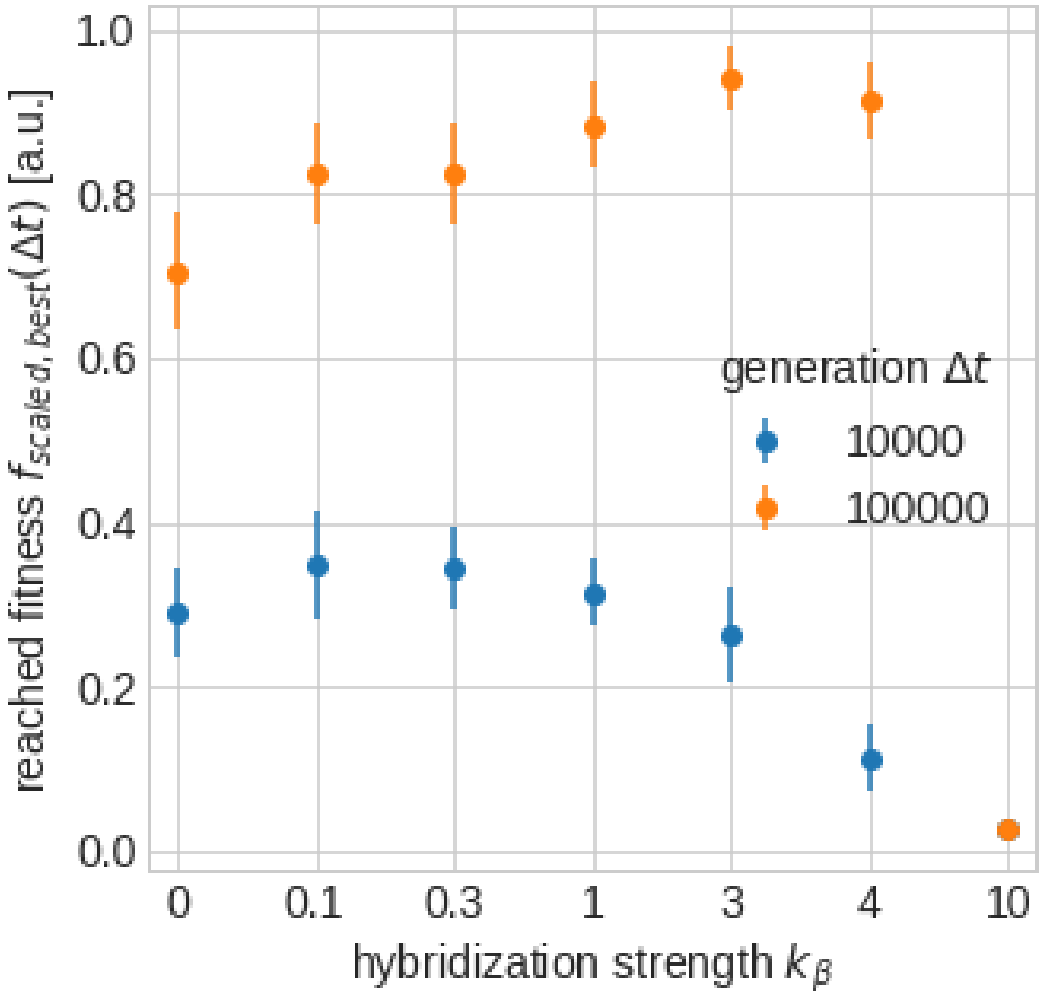


**Fig. 1.** Effect of hybridization on the efficiency of evolution at a moderate mutation rate ($p = 0.01$). Evolution rate is measured in reached fitness $f_{\text{scaled,best}}$ of the best individual under the presence of hybridization $k_\beta > 0$ compared to no hybridization $k_\beta = 0$. Even in early generations $\Delta t = 10000$ (blue dots) the evolved fitness in populations with a low hybridization rate $k_\beta <= 1$ is at a comparable height to that of populations without hybridization. Late generations $\Delta t = 100000$ (orange dots) reveal not only an accelerated evolution for populations with hybridization $k_\beta > 0$ but also a shift from the highest overall achieved fitness from $k_\beta = 0.3$ at generation $\Delta t = 10000$ to $k_\beta = 3$ at generation $\Delta t = 100000$. The fitness $f_{\text{scaled,best}}$ of the best sequence of a population is averaged over $25$ simulations, with error bars showing the standard error of the mean, using volume $V = 10$.

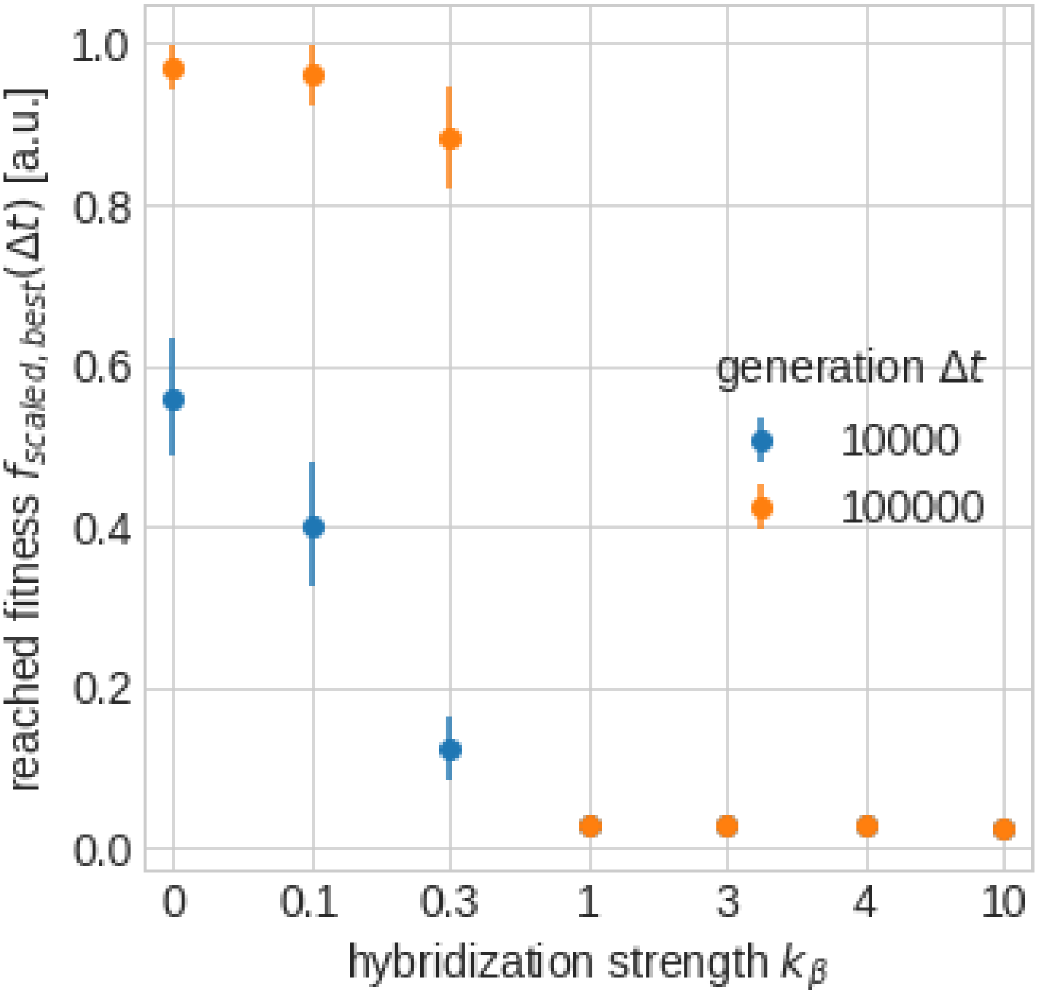


**Fig. 2.** Effect of hybridization on the efficiency of evolution at a high mutation rate close to the error threshold ($p = 0.02$). Early $\Delta t = 10000$ (blue dots) and late $\Delta t = 100000$ (orange dots) generations reveal upon increasing hybridization $k_\beta$ a decrease in efficiency of evolution. For high hybridization $k_\beta >= 1$ even no fitness improvement is observable. Only low hybridization rates $k_\beta = \{0.1, 0.3\}$ are able of reaching similar values for the fitness in late generations. Mean and standard error of the mean shown, based on $25$ simulations each, using volume $V = 10$.

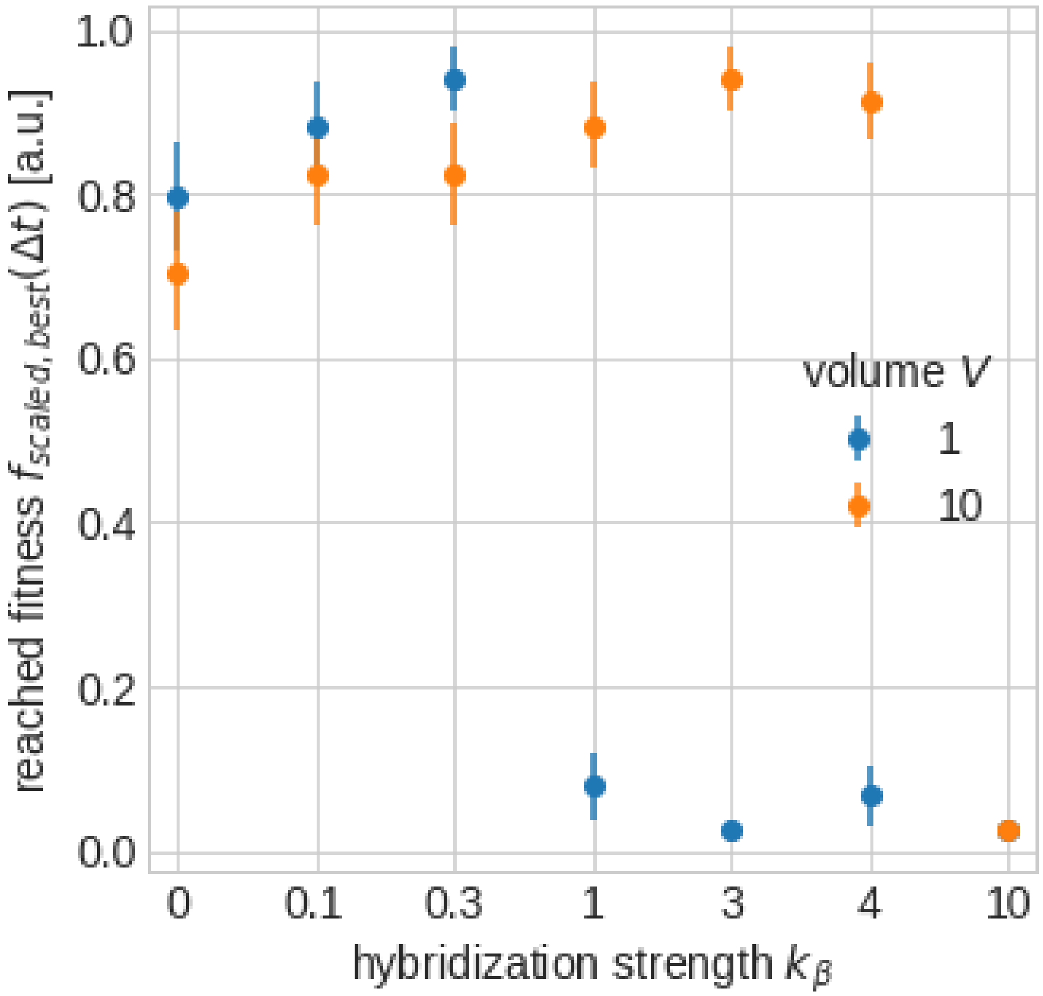


**Fig. 3.** Influence of the volume on the efficiency of evolution. A decrease in volume to $V = 1$ (blue dots) leads to an increased fitness $f_{\text{scaled,best}}$ for small hybridization rates ($k_\beta < 1$). However, for large hybridization rates ($k_\beta \geq 1$) it has a negative effect. Note that a decrease in volume is equivalent to an increase of the second order reaction rates (replication and hybridization rates, here). Mean and standard error of the mean shown, based on $25$ simulations each, using volume $V = 10$, a moderate mutation rate $p = 0.01$, and $\Delta t = 100000$.

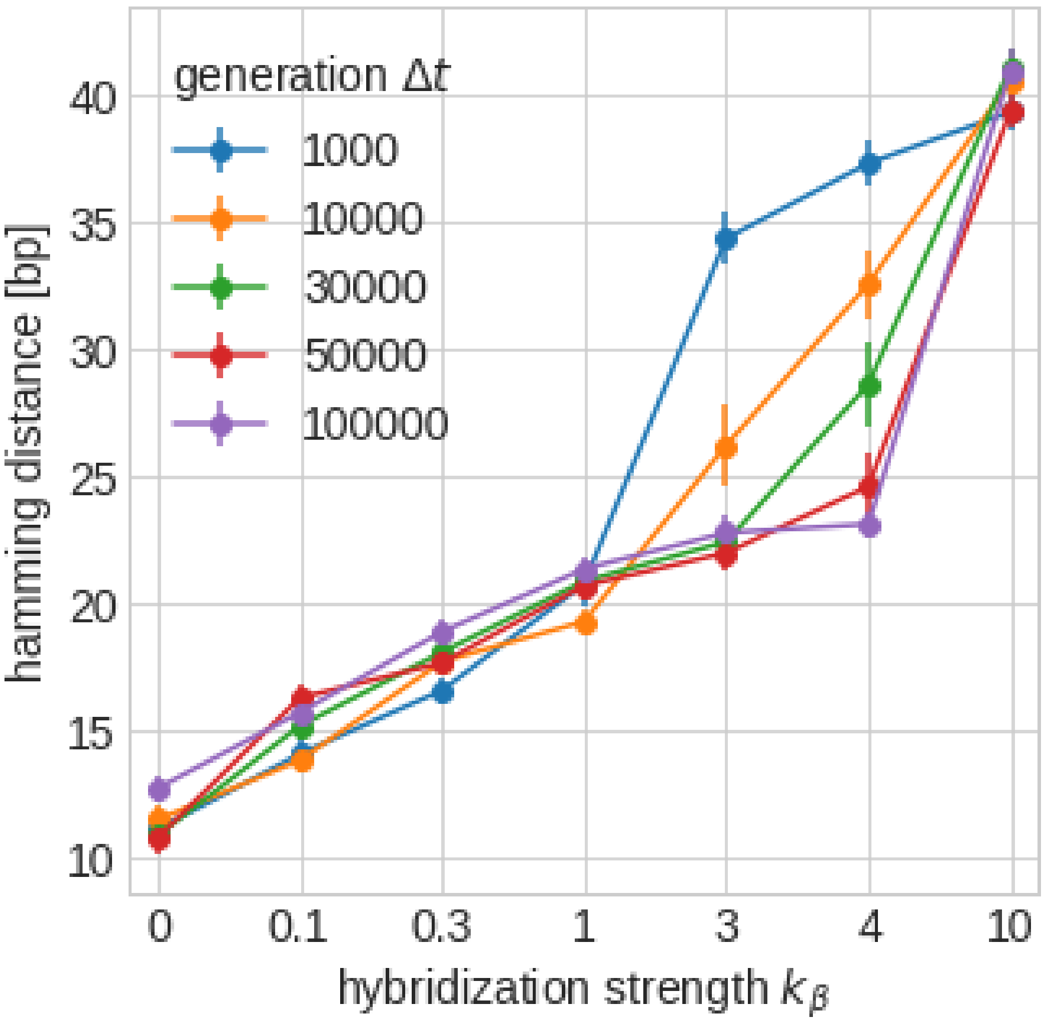
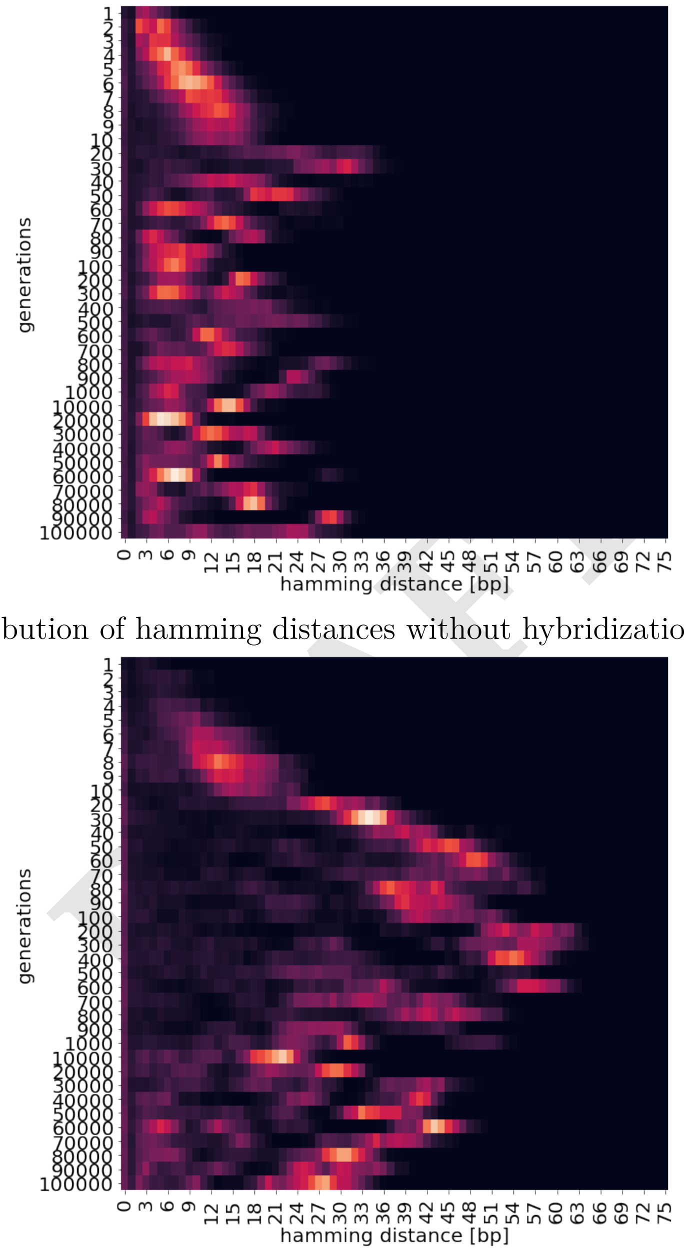


**Fig. 4.** Average hamming distance between two sequences in a population depending on the hybridization strength $k_\beta$ and generation $\Delta t$. The distance is increasing with the hybridization strength $k_\beta$, even at small generations $\Delta t = 1000$ (blue dots). Note that the decrease of the hamming distance for hybridization strength $k_\beta = \{3, 4\}$ for time $\Delta t \geq 1000$ is preceded by an increase of the distance (cf. Figure 5(b)). For clarity $\Delta t < 1000$ not shown, here. Mean and standard error of the mean shown, based on $25$ simulations each, using volume $V = 10$, and a moderate mutation rate $p = 0.01$.



(a) Distribution of hamming distances without hybridization, $k_\beta = 0$.

(b) Distribution of hamming distances with hybridization, $k_\beta = 3$.

**Fig. 5.** Comparison of the temporal progression of the hamming distance distribution between a population without (a) and a population with hybridization (b) considering a volume $V = 10$ and a moderate mutation rate $p = 0.01$.

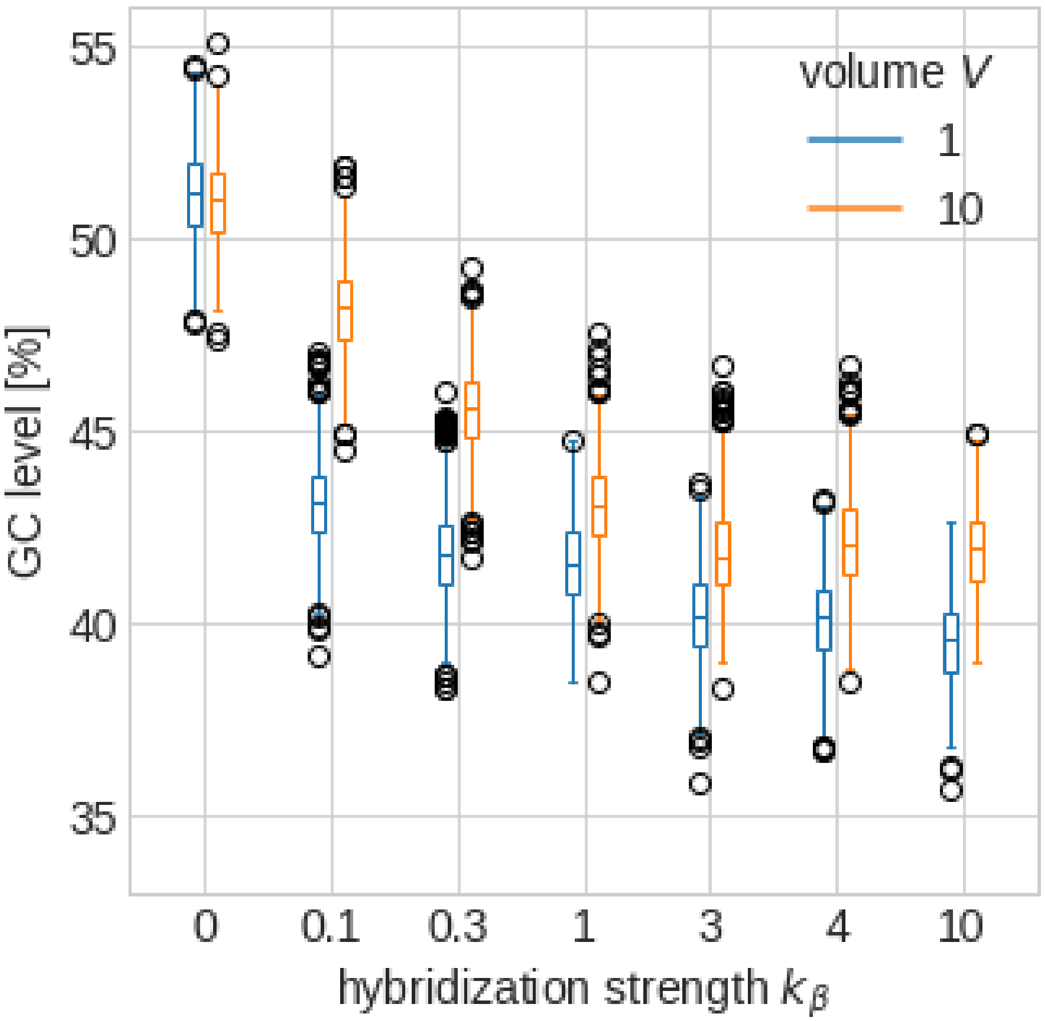


**Fig. 6.** Influence of hybridization on the GC content depending selection of sequences. An increased hybridization strength increases the selection pressure on the GC content, leading to a decrease in GC level. Depicted data considers the range of GC level of the sequence with highest fitness of each simulation, with volume $V = 10$ and mutation rate $p = 0.01$, from generation $\Delta t = 1000$ up to generation $\Delta t = 100000$. Coloured boxes contain quartiles 1 to 3 of the range of the GC level with their whiskers having 1.5 times the length of the box.

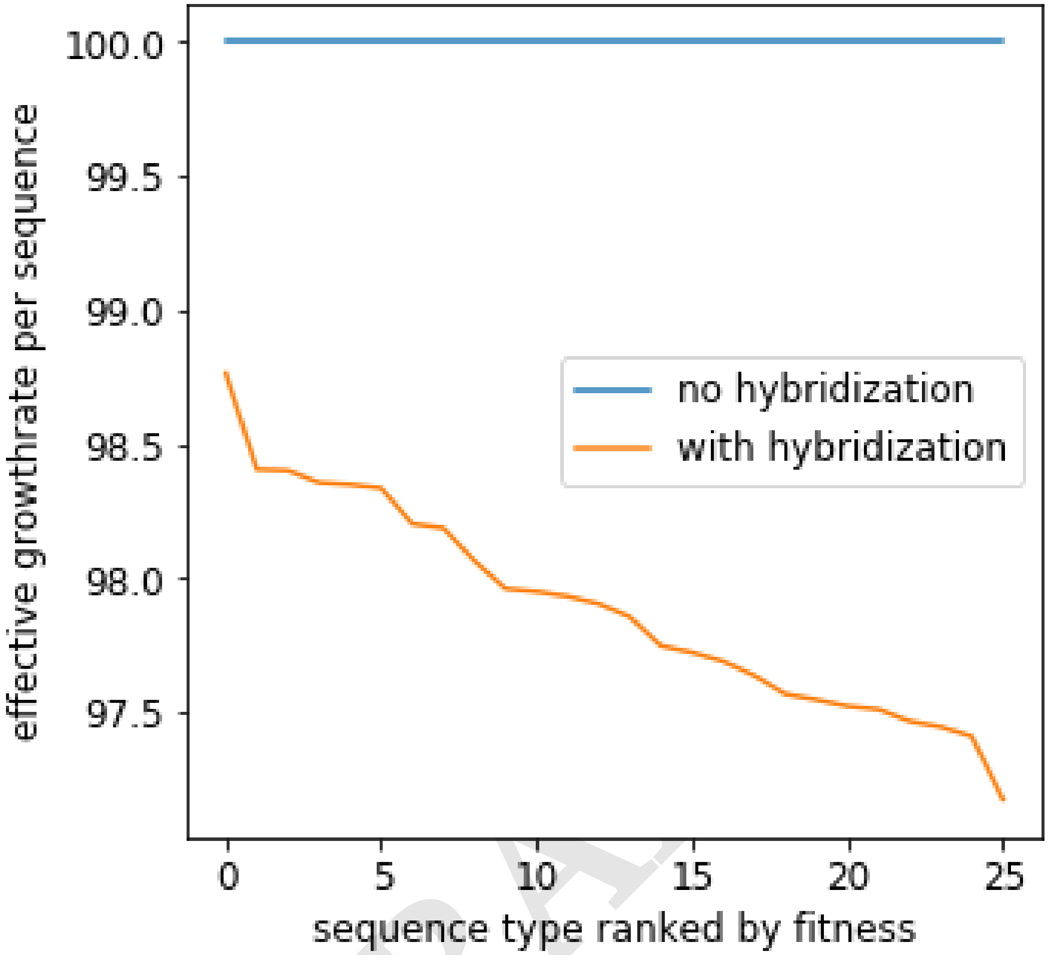


**Fig. 7.** Modulation of the effective fitness landscape (effective growth rate) due to hybridization. Neutral networks vanish under the presence of hybridization events as there are no longer different sequence types with the same effective growth rate. Depicted data shows the first 26 sequence types ranked by the effective growth rate in the final populations obtained in Figure 5, with volume $V = 10$, mutation rate $p = 0.01$ and $k_\beta \in \{0, 3\}$.

**Hybridization adds an additional selection pressure on a sequence's GC content.** Surprisingly the minimal difference of parameters of nearest neighbor thermodynamics which are the basis for the chances of a hybridization reaction expands selection to the sequence level while pure replication only selects for secondary structure (Fig. 6). The observed selection pressure towards lower GC-content in order to avoid hybridization increases with rising chances of a hybridization event but scales down with increasing volume making it harder for sequences to collide. Note that the sequence of our target structure has a high GC-content. So, this selection pressure does not push evolution towards this sequence.

**Hybridization dynamically modulates the fitness landscape.** Effective fitness of a sequence $r$ with respect to a population $P$ is defined as its effective growth rate. It is not only determined by its replication and decay rate but also by the effect of hybridization with other sequences of the population. Thus, the evolving population dynamically modulates its effective fitness landscape through hybridization, while in a population without hybridization the first 26 sequence types show the same effective growth rate. In a population with hybridization each sequence type has in general

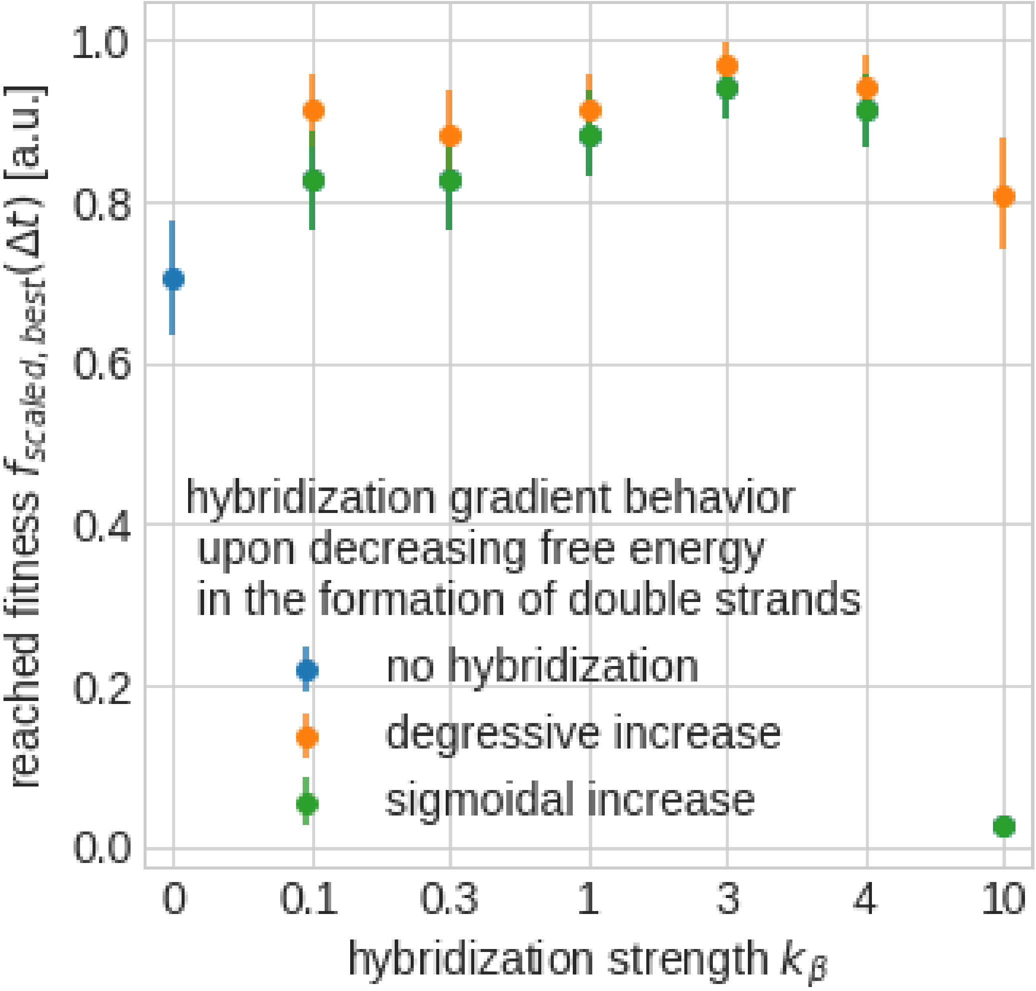


**Fig. 8.** Comparison of two different hybridization models namely the general sigmoidal (green dots) and the degressive (yellow dots) scaling of the hybridization coefficient. Only relatively high ($k_\beta = 10$) hybridization rates reveal a significant difference between the two models indicating a greater evolutionary benefit for a degressive increase of the hybridization coefficient. Mean and standard error of the mean shown, based on $25$ simulations each, using volume $V = 10$, and a moderate mutation rate $p = 0.01$.

a different effective growth rate (Fig. 7). As a consequence, there are no neutral networks anymore because the change in number of one sequence $r$ leads to an individual change in effective fitness in all sequences depending on their general probability of hybridizing with that sequence $r$.

**The observed effects are robust with respect to the chosen hybridization model.** The effect of improved evolution rate through hybridization has been observed using a hybridization coefficient based on the Hamming distance instead of the nearest-neighbor thermodynamics of oligonucleotides (34) (data not shown). Furthermore, we varied the scaling. In the results presented so far we applied a sigmoidal scaling for mapping the Gibbs free energy to stochastic rate constants for hybridization (Eq. 8), allowing for a sharp turning between high and low chances for a hybridization event upon colliding sequences. But even for degressive scaling

$$h'_{\text{scaled}}(r, r') = \frac{h(r, r')}{0.55 + h(r, r')} \qquad [9]$$

we can observe a similar accelerated evolution (Fig. 8).

## Discussion

A novel and proven exact extension of the Gillespie stochastic simulation approached allowed us to perform exact stochastic simulations of the evolution of replicating and hybridizing RNA sequences [†]. Simulations showed that hybridization, though reducing the rate of replication, can increase the rate of evolution, measured as fitness gain within a period of time.

The positive effect of hybridization has been observed for a mutation rate $p$ that has a certain distance to the error-threshold. For such a "safe" mutation rate, hybridization has lead to an improvement of replication rate and thus fitness over the long run, e.g., 10000 generations. With hybridization populations are able to expand more freely in sequence space, providing more opportunities for a benefiting mutation in terms of a higher replication rate. The effect is closely related to the limiting similarity principle(35), saying that phenotype difference between two species on the scale of the competition width is required for coexistence(36).

However, we have observed the highest rate of evolution for no hybridization and mutation rate $p$ being close to the error threshold. On the other hand a lower $p$ associated with a more conserved sequence is in favor to maintain an even longer sequence pattern. Hybridization might be of use to limit the exponential growth of sequences with that pattern such that other sequences without that particular pattern can still exist and posses the possibility in finding an even better pattern for replication.

But is there a scenario where a lower $p$ is preferable and hybridization would be useful to increase the rate of evolution? RNA viruses though replicating with relatively high error compared to their hosts (29) have a potential in the observed beneficial effect of hybridization. Using our model, further investigations may reveal the actual impact of hybridization in natural evolution.

## Materials and Methods

**Hybridization Coefficient** $h(r, r')$**.** For a more realistic model of hybridization probability upon colliding sequences $r$ and $r'$ instead of pure sequence similarity the change of Gibbs free energy upon dimer formation based on nearest-neighbor thermodynamics of oligonucleotides is used (34).

$$h(r,r') = \frac{\sum_{i\in R(r,r')} \Delta G^\circ(i)}{\Delta G^\circ(((GC)^{l/2},(GC)^{l/2}))}, \qquad [10]$$

$$\begin{aligned}\Delta G^\circ(i) &= \min(0, \sum_{j=1}^{10} n_{ij}\Delta G^\circ(j) + x_i \Delta G^\circ_{\text{init w/term } G\cdot C} + \\ &\quad (2-x_i)\Delta G^\circ_{\text{init w/ term } A\cdot T}),\end{aligned} \qquad [11]$$

$$\begin{aligned}R(r,r') &= \{(r_{k-m}, r'_{k-m}) : k,m \in \mathbb{N}, 0 < k < m < l, \\ &\quad d(r_{k-m}, r'_{k-m} = 0, r_{k-1} \neq r'_{k-1}, \\ &\quad r_{m+1} \neq r'_{m+1}\}.\end{aligned} \qquad [12]$$

The hybridization coefficient $h(r, r')$ itself is thereby the normalized sum of the change of Gibbs free energy of the sequence pair $(r, r')$ compared to the maximal change of Gibbs free energy via the sequence pair $((GC)^{l/2}, (GC^{l/2}))$ with $l$ being the length of the sequences $r$ and $r'$. Only matching regions $R(r, r')$ contribute to hybridization through the possible formation of hydrogen bonds. As only substitution

[†] The code of the complete model is at https://git.uni-jena.de/ne78xoy/hr-sim-newgillespie.

mutations take place the chance of improving matching regions by insertion of gaps is very small such that it can be omitted. The change of Gibbs free energy in one region $\Delta G^{\circ}(i)$ is dependent on the number of flanking $G \cdot C$ pairs $x_i$ and the numbers of possible base sequences $n_{ij}$ with their change of free energy parameter is taken correspondingly from table 1 (34).

| j | sequence | parameter, kcal/mol |
|---|---|---|
| 1 | AA/TT | -1.00 |
| 2 | AT/TA | -0.88 |
| 3 | TA/AT | -0.58 |
| 4 | CA/GT | -1.45 |
| 5 | GT/CA | -1.44 |
| 6 | CT/GA | -1.28 |
| 7 | GA/CT | -1.30 |
| 8 | CG/GC | -2.17 |
| 9 | GC/CG | -2.24 |
| 10 | CC/GG | -1.84 |
|  | Init. w/term. $G \cdot C$ | 0.98 |
|  | Init. w/term. $A \cdot T$ | 1.03 |

**Table 1. *Unified NN-parameters* as in (34). Parameter is the energy difference of the given basepair combination upon hybridization. Basepair combination VW/XY corresponds to sequence sections VW and YX in $3'$-to-$5'$ orientation. With respect to RNA T is substituted by U for the present model.**

**Target Structure.** Like Fontana(27) and Kupczok&Dittrich (28) we used a shape of a tRNA as the target secondary structure obtained from the sequence
$r_{target} = "GGGCAGAUAGGGCGUGUGAUAGCCCAUAGCGAACCCCCCGCUGAG$
$CUUGUGCGACGUUUGUGCACCCUGUCCCGCU"$
giving
$fold(r_{target}) =$ ((((((...((((........)))).(((((.......))))).....((((.((....)))))).)))))).... .

**Stochastic Simulation.** The new algorithm to get stochastic correct trajectories is derived from the Gillespie algorithm (16) and can be found in the supplemental text, Section S1. The major difference arises from the separation of selecting the reaction type and selecting the explicit educts. This allows the formulation of implicit chemical reactions such that not every sequence needs a separate set of replication, hybridization, and decay reactions.

**Lemma 1.** *The developed algorithm is equivalent to the Gillespie algorithm by generating a statistically correct trajectory.*

*Proof.* Both algorithms are equivalent if the following criteria are met:

1. The probability of choosing a certain sequence is the same in both variants.
2. The probability of choosing a certain reaction is the same in both variants.
3. The time interval between two successive reactions is the same in both variants.

□

These three criteria are proven in the supplemental text, Section S2.

**ACKNOWLEDGMENTS.** We acknowledge financial support by the Ministry for Economics, Sciences and Digital Society of Thuringia (TMWWDG), under the framework of the Landesprogramm ProDigital (DigLeben-5575/10-9).

1. JE Jones, V Le Sage, SS Lakdawala, Viral and host heterogeneity and their effects on the viral life cycle. *Nat. Rev. Microbiol.* (2020).
2. W Gilbert, Origin of life: The RNA world. *Nature* **319**, 618–618 (1986).
3. P Adamski, et al., From self-replication to replicator systems en route to de novo life. *Nat. Rev. Chem.* **4**, 386–403 (2020).
4. KL Vay, LI Weise, K Libicher, J Mascarenhas, H Mutschler, Templated self-replication in biomimetic systems. *Adv. Biosyst.* **3**, 1800313 (2019).
5. I Haruna, S Spiegelman, Autocatalytic synthesis of a viral RNA in vitro. *Science* **150**, 884–886 (1965).
6. G von Kiedrowski, A self-replicating hexadeoxynucleotide. *Angewandte Chemie Int. Ed. Engl.* **25**, 932–935 (1986).
7. WS Zielenski, LE Orgel, Autocatalytic synthesis of a tetranucleotide analogue. *Nature* **327**, 346–347 (1987).
8. B Liu, et al., Spontaneous emergence of self-replicating molecules containing nucleobases and amino acids. *J. Am. Chem. Soc.* **142**, 4184–4192 (2020).
9. A Salditt, et al., Thermal habitat for rna amplification and accumulation. *Phys. Rev. Lett.* **125**, 048104 (2020).
10. P van Nies, et al., Self-replication of dna by its encoded proteins in liposome-based synthetic cells. *Nat. Commun.* **9** (2018).
11. KB Mullis, The unusual origin of the polymerase chain reaction. *Sci. Am.* **262**, 56–65 (1990).
12. E Szathmáry, Simple growth laws and selection consequences. *Trends Ecol. & Evol.* **6**, 366–370 (1991).
13. CK Biebricher, M Eigen, WC Gardiner Jr, Kinetics of RNA replication: competition and selection among self-replicating rna species. *Biochemistry* **24**, 6550–6560 (1985).
14. M Eigen, P Schuster, A principle of natural self-organization. *Naturwissenschaften* **64**, 541–565 (1977).
15. N Rohde, H Daum, CK Biebricher, The mutant distribution of an RNA species replicated by Q$\beta$ replicase. *J. Mol. Biol.* **249**, 754–762 (1995).
16. DT Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**, 403–434 (1976).
17. G Meszéna, E Szathmáry, Adaptive dynamics of parabolic replicators. *Selection* **2**, 147–159 (2002).
18. P Dittrich, Ph.D. thesis (University of Dortmund, Department of Computer Science,D-44221 Dortmund, Germany) (2001).
19. HC Ito, U Dieckmann, A new mechanism for recurrent adaptive radiations. *The Am. Nat.* **170**, E96–E111 (2007).
20. HC Ito, U Dieckmann, Evolutionary branching under slow directional evolution. *J. Theor. Biol.* **360**, 290–314 (2014).
21. A Szilágyi, et al., Review ecology and evolution in the rna world dynamics and stability of prebiotic replicator systems. *Life* **7**, 48 (2017).
22. E Szathmáry, JM Smith, From replicators to reproducers: the first major transitions leading to life. *J. Theor. Biol.* **187**, 555–571 (1997).
23. R Mizuuchi, K Usui, N Ichihashi, Structural transition of replicable rnas during in vitro evolution with q$\beta$ replicase. *RNA* **26**, 83–90 (2020).
24. W Fontana, P Schuster, A computer model of evolutionary optimization. *Biophys. Chem.* **26**, 123–147 (1987).
25. R Nussinov, G Pieczenik, JR Griggs, DJ Kleitman, Algorithms for loop matchings. *SIAM J. on Appl. Math.* **35**, 68–82 (1978).
26. C Reidys, PF Stadler, P Schuster, Generic properties of combinatory maps: neutral networks of rna secondary structures. *Bull. Math. Biol.* **59**, 339–397 (1997).
27. W Fontana, P Schuster, Continuity in evolution: on the nature of transitions. *Science* **280**, 1451–1455 (1998).
28. A Kupczok, P Dittrich, Determinants of simulated rna evolution. *J. Theor. Biol.* **238**, 726–735 (2006).
29. SDW Frost, BR Magalis, SL Kosakovsky Pond, Neutral theory and rapidly evolving viral pathogens. *Mol. Biol. Evol.* **35**, 1348–1354 (2018).
30. AS Lauring, Within-host viral diversity: A window into viral evolution. *Annu. Rev. Virol.* **7**, 63–81 (2020).
31. IR Epstein, Competitive coexistence of self-reproducing macromolecules. *J. Theor. Biol.* **78**, 271–298 (1979).
32. R Lorenz, et al., ViennaRNA package 2.0. *Algorithms for Mol. Biol.* **6**, 26 (2011).
33. M Eigen, Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58**, 465–523 (1971).
34. JJ SantaLucia, A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA* **95**, 1460–1465 (1998).
35. R MacArthur, R Levins, The limiting similarity, convergence, and divergence of coexisting species. *The Am. Nat.* **101**, 377–385 (1967).
36. P Szabó, G Meszéna, Limiting similarity revisited. *Oikos* **112**, 612–619 (2006).

# Product inhibition can accelerate evolution - Supplemental Text

**Beatrice Ruth**[a] **and Peter Dittrich**[a,b]

[a]Department of Mathematics and Computer Science, Friedrich Schiller University Jena, 07744 Jena, Germany; [b]Jena Centre for Bioinformatics, 07743 Jena, Germany

## S1 Simulation algorithm

1. **Initialization:**

   (a) Set the time variable $t = 0$.

   (b) Specify and store the implicit molecule categories $S_{\text{sequence}}$ and $S_{\text{substrate}}$ with their numbers $n_{\text{sequence}}$ and $n_{\text{substrate}}$.

   (c) Specify and store the split into the explicit molecule categories $S_{\text{sequence},i}$ and $S_{\text{substrate},s}$ with their numbers $n_{\text{sequence},1}, n_{\text{sequence},2}, \ldots, n_{\text{sequence},L_{\text{sequence}}}$ and $n_{\text{substrate},s}$ (In this case: $n_{\text{substrate},s} = 100$). The total number of sequences is computable by $n_{\text{sequence}} = \sum_{i=1}^{L_{\text{sequence}}} n_{\text{sequence},i}$.

   (d) Specify and store the implicit chemical reactions $R_\mu$ with $\mu \in \{replication, hybridization, decay\}$ as $S_{\text{sequence}} + S_{\text{substrate}} \rightarrow 2S_{\text{sequence}}$ (replication), $2S_{\text{sequence}} \rightarrow 2S_{\text{substrate}}$ (hybridization) and $S_{\text{sequence}} \rightarrow S_{\text{substrate}}$ (decay).

   (e) Calculate and store the quantities $c_{\text{replication}} = k_{\text{replication}}/V, c_{\text{hybridization}} = k_{\text{hybridization}}/V, c_{\text{decay}} = k_{\text{decay}}$ ($k_\mu$ corresponds to the rate constant and $V$ is the volume).

   (f) Specify and store the calculation rules $f_{\text{replication}}(S_{\text{sequence},i}) = \frac{0.01}{1.01 - f(S_{\text{sequence},i})}$ with $f(S_{\text{sequence},i})$ representing the fitness of the explicit sequence $i$,
   $f_{\text{hybridization}}(S_{\text{sequence},i}, S_{\text{sequence},j}) = \frac{h(S_{\text{sequence},i}, S_{\text{sequence},j})^p}{w^p + h(S_{\text{sequence},i}, S_{\text{sequence},j})^p}$ with $h(S_{sequence,i}, S_{sequence,j})$ representing the hybridization coefficient of the explicit sequences $i$ and $j$, $w$ and $p$ correspond to the turning and exponent of the sigmoid function, and $f_{decay}(S_{sequence,i}) = 1$ to model the explicit chemical reactions.

   (g) Store the maximal achievable values of those calculation rules $e_{replication} = 1$, $e_{hybridization} = 1$ and $e_{decay} = 1$.

   (h) Calculate and store all 3 reaction propensities $a_{replication} = c_{replication} n_{sequence} n_{substrate} e_{replication}$, $a_{hybridization} = c_{hybridization} n_{sequence} \frac{n_{sequence} - 1}{2} \cdot e_{hybridization}$ and $a_{decay} = c_{decay} n_{sequence}$.

   (i) Specify and store a series of sampling times $t_1 < t_2 < \cdots$ , and also a stopping time $t_{stop}$. (1)

2. **Timestep:** Generate a random number $r_1$ uniformly distributed on $[0, 1]$ to determine the time interval $\tau = \frac{1}{a0_{max}} log(\frac{1}{r_1})$ with $a0_{max} = a_{replication} + a_{hybridization} + a_{decay}$.

3. **Reaction type:**

   (a) Calculate the contribution of $a_{\text{replication}}$, $a_{\text{hybridization}}$ and $a_{\text{decay}}$ to their sum as $\eta_\mu = \frac{a_\mu}{a_{\text{replication}} + a_{\text{hybridization}} + a_{\text{decay}}}$ with $\mu \in \{\text{replication, hybridization, decay}\}$.

   (b) Proportional selection of the reaction type $R_\mu$ based on $\eta_\mu$.

4. **Reaction probability:**

   (a) Proportional selection of the required sequence/s based on their relative frequency $\frac{n_{\text{sequence, i}}}{n_{\text{sequence}}}$.

   (b) Calculate the probability $p_\mu$ for the chosen reaction type $R_\mu$ taking into account the selected sequence/s $S_{\text{sequence, i}}$, $S_{\text{sequence, j}}$ via $p_\mu = \frac{f_\mu(S_{\text{sequence, i}})}{e_\mu}$ if $\mu \in \{replication, decay\}$ else $p_\mu = \frac{f_\mu(S_{sequence,i}, S_{sequence,j})}{e_\mu}$.

5. **Update:**

   (a) Advance time $t$ by the random generated time step $t = t + \tau$ in step 2 time step.

   (b) Generate a random number $r_2$ uniformly distributed on $[0, 1]$. If $r_2 < p$ remove all educts and add all products of $R_\mu$ else proceed with the next step iteration.

(c) If replication takes place add once the removed sequence and generate a mutated sequence from the explicit sequence. If the mutated sequence is already present add 1 to $n_{sequence,i}$ of the corresponding sequence $S_{sequence,i}$ else $L_{sequence} = L_{sequence} + 1$ and $n_{sequence,L_{sequence}} = 1$ with $S_{sequence,L_{sequence}}$ being the newly mutated sequence.

(d) Delete all $S_{sequence,i}$ with $n_{sequence,i} = 0$. For each deleted $S_{sequence,i}$ adjust the following indices $i + x$ with $x \in \mathbb{N}*$ by $i + x - 1$ and $L_{sequence} = L_{sequence} - 1$.

(e) Recalculate $a_{replication} = c_{replication} n_{sequence} n_{substrate} e_{replication}$, $a_{hybridization} = c_{hybridization} n_{sequence} \frac{n_{sequence}-1}{2} \cdot e_{hybridization}$ and $a_{decay} = c_{decay} n_{sequence}$.

6. **Iteration:**

(a) If $t$ has just been advanced through one of the sampling times $t_i$, read out the current molecular population values $S_{sequence,1}, S_{sequence,2}, \dots, S_{sequence,L_{sequence}}$.

(b) If $t > t_{stop}$, or if no more reactions are possible (all $a_\mu = 0$), terminate the calculation; otherwise, return to Step 2 time step.

## S2 Proof of Lemma 1

For proving Lemma 1 from the main text

**Lemma 1.** *The developed algorithm is equivalent to the Gillespie algorithm by generating a statistically correct trajectory.*

we need to prove the following three sub-lemmata:

**Lemma 1.1.** *The probability of choosing a certain sequence is the same in both variants.*

*Proof.* Through the rate equations the following is obtained for the Gillespie algorithm:

$$a_{\text{replication}}(S_{\text{sequence},i}) = \frac{k_\alpha}{V} f_{\text{scaled}}(S_{\text{sequence},i}) n_{\text{sequence},i} n_{\text{substrate}}, \quad [1]$$

$$a_{\text{hybridization}}(S_{\text{sequence},i}, S_{\text{sequence},i}) = \frac{k_\beta}{V} h_{\text{scaled}}(S_{\text{sequence},i}, S_{\text{sequence},i}) n_{\text{sequence},i} \frac{n_{\text{sequence},i} - 1}{2}, \quad [2]$$

$$a_{\text{hybridization}}(i,j) = \frac{k_\beta}{V} h_{\text{scaled}}(S_{\text{sequence},i}, S_{\text{sequence},j}) n_{\text{sequence},i} n_{\text{sequence},j}, i \neq j, \quad [3]$$

$$a_{\text{decay}}(S_{\text{sequence},i}) = k_\gamma n_{\text{sequence},i}. \quad [4]$$

And for the developed algorithm:

$$a_{\text{replication}} = \frac{k_\alpha}{V} f_{\text{scaled, max}} n_{\text{sequence}} n_{\text{substrate}}, \qquad n_{\text{sequence}} = \sum_{i=1}^{L_{\text{sequence}}} n_{\text{sequence},i}, \quad [5]$$

$$a_{\text{hybridization}} = \frac{k_\beta}{V} h_{\text{scaled, max}} n_{\text{sequence}} \frac{n_{\text{sequence}} - 1}{2}, \quad [6]$$

$$a_{\text{decay}} = k_\gamma n_{\text{sequence}}. \quad [7]$$

in the developed algorithm only after choosing the reaction type the explicit molecules in this case sequences are randomly chosen based on their relative abundancies reacting accordingly to their specific reaction parameter ($f_{\text{scaled}}(S_{\text{sequence},i})$, $h_{\text{scaled}}(S_{\text{sequence},i}, S_{\text{sequence},j})$), leading to:

$$a_{\text{replication}}(S_{\text{sequence},i}) = \frac{k_\alpha}{V} f_{\text{scaled,max}} n_{\text{sequence}} n_{\text{substrate}} \frac{n_{\text{sequence},i}}{n_{\text{sequence}}} \cdot \frac{f_{\text{scaled}}(S_{\text{sequence},i})}{f_{\text{scaled,max}}}, \quad [8]$$

$$a_{\text{hybridization}}(S_{\text{sequence},i}, S_{\text{sequence},i}) = \frac{k_\beta}{V} h_{\text{scaled,max}} n_{\text{sequence}} \frac{n_{\text{sequence}} - 1}{2} \frac{n_{\text{sequence},i}}{n_{\text{sequence}}} \cdot \frac{n_{\text{sequence},i} - 1}{n_{\text{sequence}} - 1} \frac{h_{\text{scaled}}(S_{\text{sequence},i}, S_{\text{sequence},i})}{h_{\text{scaled,max}}}, \quad [9]$$

$$a_{\text{hybridization}}(i,j) = \frac{k_\beta}{V} h_{\text{scaled,max}} n_{\text{sequence}} \frac{n_{\text{sequence}} - 1}{2} \cdot \frac{n_{\text{sequence},i} n_{\text{sequence},j} + n_{\text{sequence},j} n_{\text{sequence},i}}{n_{\text{sequence}}(n_{\text{sequence}} - 1)} \cdot \frac{h_{\text{scaled}}(S_{\text{sequence},i}, S_{\text{sequence},j})}{h_{\text{scaled,max}}}, i \neq j, \quad [10]$$

$$a_{\text{decay}}(S_{\text{sequence},i}) = k_\gamma n_{\text{sequence}} \frac{n_{\text{sequence},i}}{n_{\text{sequence}}}. \quad [11]$$

Since $n_{\text{sequence}}$, $f_{\text{scaled,max}}$ and $h_{\text{scaled,max}}$ in the equations above cancel each other out the equations of the Gillespie algorithm are obtained. Thus leading to the conclusion that in both cases a certain sequence has the same probability to undergo a reaction. □

**Lemma 1.2.** *The probability of choosing a certain reaction is the same in both variants.*

*Proof.* The selection of a reaction type in the Gillespie algorithm:

$$\begin{aligned}
a0 &= \sum_{i=1}^{L_{\text{sequence}}} a_{\text{replication}}(S_{\text{sequence},i}) \\
&+ \sum_{i=1}^{L_{\text{sequence}}} a_{\text{decay}}(S_{\text{sequence},i}) \\
&+ \sum_{i=1}^{L_{\text{sequence}}} \sum_{j=1}^{i} a_{\text{hybridization}}(S_{\text{sequence},i}, S_{\text{sequence},j}),
\end{aligned} \tag*{[12]}$$

$$\sum_{i=1}^{L_{\text{sequence}}} a_{\text{replication}}(S_{\text{sequence},i}) = \sum_{i=1}^{L_{\text{sequence}}} \frac{k_\alpha}{V} f_{\text{scaled}}(S_{\text{sequence},i}) n_{\text{sequence},i} n_{\text{substrate}}, \tag*{[13]}$$

$$\begin{aligned}
\sum_{i=1}^{L_{\text{sequence}}} a_{\text{hybridization}}(i,i) &= \sum_{i=1}^{L_{\text{sequence}}} \frac{k_\beta}{V} h_{\text{scaled}}(S_{\text{sequence},i}, S_{\text{sequence},i}) \\
&\quad n_{\text{sequence},i} \frac{n_{\text{sequence},i} - 1}{2},
\end{aligned} \tag*{[14]}$$

$$\begin{aligned}
\sum_{i=2}^{L_{\text{sequence}}} \sum_{j=1}^{i-1} a_{\text{hybridization}}(S_{\text{sequence},i}, S_{\text{sequence},j}) &= \sum_{i=2}^{L_{\text{sequence}}} \sum_{j=1}^{i-1} \frac{k_\beta}{V} h_{\text{scaled}}(S_{\text{sequence},i}, S_{\text{sequence},j}) \\
&\quad n_{\text{sequence},i} n_{\text{sequence},j},
\end{aligned} \tag*{[15]}$$

$$\sum_{i=1}^{L_{\text{sequence}}} a_{\text{decay}}(S_{\text{sequence},i}) = \sum_{i=1}^{L_{\text{sequence}}} k_\gamma n_{\text{sequence},i} = k_\gamma n_{\text{sequence}} \equiv a_{\text{decay}}. \tag*{[16]}$$

On the other hand the selection of a reaction type in the developed algorithm:

$$a0_{\text{max}} = a_{replication} + a_{hybridization} + a_{decay}. \tag*{[17]}$$

Reshaping the equation of the replication sum in the Gillespie algorithm:

$$\begin{aligned}
\sum_{i=1}^{L_{sequence}} a_{replication}(S_{sequence,i}) &= \sum_{i=1}^{L_{sequence}} \frac{k_\alpha}{V} (f_{scaled,max} - d_f(S_{sequence,i})) n_{sequence,i} n_{substrate}, \\
&\quad f_{scaled}(S_{sequence,i}) = f_{scaled,max} - d_f(S_{sequence,i}), \\
&= \sum_{i=1}^{L_{sequence}} \frac{k_\alpha}{V} f_{scaled,max} n_{sequence,i} n_{substrate} \\
&\quad - \sum_{i=1}^{L_{sequence}} \frac{k_\alpha}{V} d_f(S_{sequence,i}) n_{sequence,i} n_{substrate}, \\
&= \frac{k_\alpha}{V} f_{scaled,max} n_{sequence} n_{substrate} - \sum_{i=1}^{L_{sequence}} \frac{k_\alpha}{V} d_f(S_{sequence,i}) n_{sequence,i} n_{substrate}, \\
&= a_{replication} - \sum_{i=1}^{L_{sequence}} \frac{k_\alpha}{V} d_f(S_{sequence,i}) n_{sequence,i} n_{substrate}.
\end{aligned} \tag*{[18]}$$

Respecting the fact of denying a replication in the developed algorithm:

$$\begin{aligned}
a_{noreplication}(S_{sequence,i}) &= \frac{k_\alpha}{V} f_{scaled,max} n_{sequence,i} n_{substrate} (1 - \frac{f_{scaled}(S_{sequence,i})}{f_{scaled,max}}),\\
&= \frac{k_\alpha}{V} f_{scaled,max} n_{sequence,i} n_{substrate} \frac{f_{scaled,max} - f_{scaled}(S_{sequence,i})}{f_{scaled,max}},\\
\sum_{i=1}^{L_{sequence}} a_{noreplication}(S_{sequence,i}) &= \sum_{i=1}^{L_{sequence}} \frac{k_\alpha}{V} d_f(S_{sequence,i}) n_{sequence,i} n_{substrate}.
\end{aligned} \tag*{[19]}$$

Since the obtained equation is equal to the difference of the direct reaction selection between the two algorithms both algorithms posses the same possibility for a replication event.
Reshaping the hybridization sum for equal sequences of the Gillespie algorithm:

$$\begin{aligned}
\sum_{i=1}^{L_{sequence}} a_{hybridization}(S_{sequence,i}, S_{sequence,i}) &= \sum_{i=1}^{L_{sequence}} \frac{k_\beta}{V} (h_{scaled,max} - d_h(S_{sequence,i}, S_{sequence,i}) \cdot\\
&\quad n_{sequence,i} \frac{n_{sequence,i} - 1}{2},\\
&\quad h_{scaled}(S_{sequence,i}, S_{sequence,i}) = h_{scaled,max} - d_h(S_{sequence,i}, S_{sequence,i}),\\
&= \sum_{i=1}^{L_{sequence}} \frac{k_\beta}{V} h_{scaled,max} n_{sequence,i} \frac{n_{sequence,i} - 1}{2} -\\
&\quad \sum_{i=1}^{l} \frac{k_\beta}{V} d_h(S_{sequence,i}, S_{sequence,i}) n_{sequence,i} \frac{n_{sequence,i} - 1}{2},\\
&= \frac{k_\beta}{V} h_{scaled,max} \frac{\sum_{i=1}^{L_{sequence}} n^2_{sequence,i} - n_{sequence}}{2} -\\
&\quad \sum_{i=1}^{L_{sequence}} \frac{k_\beta}{V} d_h(S_{sequence,i}, S_{sequence,i}) n_{sequence,i} \frac{n_{sequence,i} - 1}{2}.
\end{aligned} \tag*{[20]}$$

Reshaping the hybridization sum for unequal sequences of the Gillespie algorithm:

$$\begin{aligned}
\sum_{i=2}^{L_{sequence}} \sum_{j=1}^{i-1} a_{hybridization}(S_{sequence,i}, S_{sequence,j}) &= \sum_{i=2}^{L_{sequence}} \sum_{j=1}^{i-1} \frac{k_\beta}{V} (h_{scaled,max} - d_h(S_{sequence,i}, S_{sequence,j})) \cdot\\
&\quad n_{sequence,i} n_{sequence,j},\\
&\quad h_{scaled}(S_{sequence,i}, S_{sequence,j}) = h_{scaled,max}\\
&\quad -d_h(S_{sequence,i}, S_{sequence,j}),\\
&= \sum_{i=2}^{L_{sequence}} \sum_{j=1}^{i-1} \frac{k_\beta}{V} h_{scaled,max} n_{sequence,i} n_{sequence,j} -\\
&\quad \sum_{i=2}^{L_{sequence}} \sum_{j=1}^{i-1} \frac{k_\beta}{V} d_h(S_{sequence,i}, S_{sequence,j}) \cdot\\
&\quad n_{sequence,i} n_{sequence,j}.
\end{aligned} \tag*{[21]}$$

A simplified view via adding the terms with $h_{scaled,max}$ of hybridization with equal and unequal sequences leads to:

$$\begin{aligned}
a_{hybridization,max} &= \frac{k_\beta}{V} h_{scaled,max} \frac{\sum_{i=1}^{L_{sequence}} n^2_{sequence,i} - n_{sequence}}{2} + \frac{k_\beta}{V} h_{scaled,max} \sum_{i=2}^{L_{sequence}} \sum_{j=1}^{i-1} n_{sequence,i} n_{sequence,j},\\
&= \frac{k_\beta}{V} h_{scaled,max} \frac{n^2_{sequence} - n_{sequence}}{2},\\
&\quad n^2_{sequence} = \sum_{i=1}^{L_{sequence}} n^2_{sequence,i} + 2 \sum_{i=2}^{L_{sequence}} \sum_{j=1}^{i-1} n_{sequence,i} n_{sequence,j}\\
&= \frac{k_\beta}{V} h_{scaled,max} n_{sequence} \frac{n_{sequence} - 1}{2} \equiv a_{hybridization}.
\end{aligned} \tag*{[22]}$$

The difference between both algorithms immediately after selection of hybridization is the sum of the terms containing $d_h(S_{sequence,i}, S_{sequence,j})$:

$$\begin{aligned}&\frac{k_\beta}{V}\sum_{i=1}^{L_{sequence}} d_h(S_{sequence,i}, S_{sequence,i}) n_{sequence,i}\frac{n_{sequence,i}-1}{2} + \\ &\frac{k_\beta}{V}\sum_{i=2}^{L_{sequence}}\sum_{j=1}^{i-1} d_h(S_{sequence,i}, S_{sequence,j}) n_{sequence,i} n_{sequence,j}.\end{aligned} \quad [23]$$

Respecting the fact of denying a hybridization after the selection of the explicit sequences in the developed algorithm with probability $1 - p_{hybridization}(S_{sequence,i}, S_{sequence,j}) = 1 - \frac{h_{scaled}(S_{sequence,i}, S_{sequence,j})}{h_{scaled,max}}$:

$$\begin{aligned}a_{nohybridization}(S_{sequence,i}, S_{sequence,i}) &= \frac{k_\beta}{V} h_{scaled,max} n_{sequence,i} \frac{n_{sequence,i}-1}{2} \cdot \\ &\quad (1 - \frac{h_{scaled}(S_{sequence,i}, S_{sequence,i})}{h_{scaled,max}}), \\ &= \frac{k_\beta}{V} d_h(S_{sequence,i}, S_{sequence,j}) n_{sequence,i} \cdot \\ &\quad \frac{n_{sequence,i}-1}{2},\end{aligned} \quad [24]$$

$$\begin{aligned}\sum_{i=1}^{L_{sequence}} a_{nohybridization}(S_{sequence,i}, S_{sequence,i}) &= \frac{k_\beta}{V} \sum_{i=1}^{L_{sequence}} d_h(S_{sequence,i}, S_{sequence,i}) \cdot \\ &\quad n_{sequence,i}\frac{n_{sequence,i}-1}{2}\end{aligned} \quad [25]$$

$$\begin{aligned}a_{nohybridization}(S_{sequence,i}, S_{sequence,j}) &= \frac{k_\beta}{V} h_{scaled,max} n_{sequence,i} n_{sequence,j} \cdot \\ &\quad \frac{d_h(S_{sequence,i}, S_{sequence,j})}{h_{scaled,max}}, \\ &= \frac{k_\beta}{V} d_h(S_{sequence,i}, S_{sequence,j}) \cdot \\ &\quad n_{sequence,i} n_{sequence,j},\end{aligned} \quad [26]$$

$$\begin{aligned}\sum_{i=2}^{L_{sequence}}\sum_{j=1}^{i} a_{nohybridization}(S_{sequence,i}, S_{sequence,j}) &= \frac{k_\beta}{V} \sum_{i=2}^{L_{sequence}}\sum_{j=1}^{i} d_h S_{sequence,i}, S_{sequence,j} \cdot \\ &\quad n_{sequence,i} n_{sequence,j}.\end{aligned} \quad [27]$$

The possibility of denying a hybridization event corresponds to the initial difference between both algorithms leading to an equal probability for a hybridization event. Thus, in both algorithms the probability for each reaction type is the same. □

**Lemma 1.3.** *The time interval between two successive reactions is the same in both variants.*

*Proof.* Contrary to the Gillespie algorithm where each time increment corresponds to a reaction event, in our new algorithm a time increment is possible without a reaction.
The size of the time step in the Gillespie algorithm is computed by:

$$\tau_{reaction} = \frac{1}{a0}\log(\frac{1}{r_1}). \quad [28]$$

In our algorithm we increment $m$ until a reaction happens,i.e., there are $m-1$ time steps without a reaction. Then the overall time step is:

$$\tau'_{reaction} = \frac{m}{a0_{max}}\sum_{i=1}^{m}\log(\frac{1}{r_{1,i}}), m \in \mathbb{N}^*. \quad [29]$$

where $r_{1,i}$ is the $i$-th random number drawn uniformly from $[0, 1]$. In the following we show that $\tau_{reaction}$ and $\tau'_{reaction}$ follow the same distribution.

The $m$ steps can be seen as sampling with replacement such that either a single reaction with probability $p = p_{reaction}$ takes places or no reaction with probability $1 - p$ takes place. Because replication, hybridization and decay are disjoint events, the probability $p = p_{reaction}$ is the sum of replication, hybridization and decay probabilities :

$$\begin{aligned}
p_{reaction} &= p_{replication} + p_{hybridization} + p_{decay},\\
p_{replication} &= \frac{\sum_{i=1}^{n} a_{replication}(S_{sequence,i})}{a0_{max}},\\
p_{hybridization} &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{i} a_{hybridization}(S_{sequence,i}, S_{sequence,j})}{a0_{max}},\\
p_{decay} &= \frac{\sum_{i=1}^{n} a_{decay(S_{sequence,i})}}{a0_{max}},\\
\hookrightarrow p_{reaction} &= \frac{\sum_{i=1}^{n}(a_{replication}(S_{sequence,i}) + \sum_{j=1}^{i} a_{hybridization}(S_{sequence,i}, S_{sequence,j}) + a_{decay}(S_{sequence,i}))}{a0_{max}},\\
\hookrightarrow p_{reaction} &= \frac{a0}{a0_{max}}.
\end{aligned} \tag{30}$$

Through sampling with replacement, the number of steps $m$ is distributed according to a geometric distribution with expectation $1/p$. This leads to an expected number of $m = \frac{a0_{max}}{a0}$ steps. Since a single time step $\tau$ is distributed according to an exponential distribution with expectation $\frac{1}{\lambda} = \frac{1}{a0}$ the sum of $m$ exponential distributed random variables $X$ also has to follow an exponential distribution with the same expectation. As the pdf of such a random variable $Y = \sum_{i=1}^{m} X_i$ is $f_Y(y) = \lambda p e^{-\lambda p y}$ (2), $Y$ is exponential distributed with expectation $E(Y) = \frac{1}{p\lambda}$.

$$\begin{aligned}
E(Y) &= \frac{1}{p\lambda},\\
&= \frac{1}{\frac{a0}{a0_{\mathrm{max}}} a0_{max}},\\
&= \frac{1}{a0}.
\end{aligned} \tag{31}$$

Because $E(Y)$ corresponds to $\frac{1}{a0}$, the expectation of the time interval between two reaction events in the Gillespie algorithm, it can be expected that both algorithms have time intervals of equal length between two reactions. □

1. DT Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**, 403–434 (1976).
2. Sasha, amWhy, Pdf of a sum of exponential random variables [closed] (https://math.stackexchange.com/questions/634158/pdf-of-a-sum-of-exponential-random-variables, visited: 2021-06-04) (2014).