The effect of small adapter RNAs on the evolution of a population of functional RNAs

Project by
Kóródi Lőrinc

Supervised by
dr. Paulien Hogeweg

2024/02

Introduction

One of the most important sections of the beginning of life is the RNA world (Gilbert 1986, Kun et al. 2015). In this model the life emerged from self-replicating RNA-molecules, that served as templates and catalytic agents (ribozymes). A very critical question regarding this scenario is that how could these simple molecules generate and maintain the diversity required for the functions a simple autocatalytic organism. One possible solution to this problem lies within the RNA – RNA interactions, which was investigated by De Boer and Hogeweg in 2012. In their research they looked at larger RNA molecules that co-folded with smaller adapter RNA molecules. Among other thing they found that the presence of these adapters is able to increase the number of (functional) structures an RNA can code for. Their research was done with protocells. This is where this project starts investigating a similar model but instead of protocells it was done with a population of free-floating RNA molecules on a spatial grid, that evolved towards a predetermined target secondary-structure.

Methods

Each simulation utilised the cacatoo library (van Dijk, 2022) for JavaScript. The grid had the size of 20 by 20 cells and was wrapped from each side. Each cell contained a 50 base long RNA. In each timepoint every cell had the chance of being overgrown by its neighbours. This was done synchronously and according to the following procedure. A cell can be overgrown by any of the neighbouring 8 cells, which was at least as close to the target structure as the cell. The distances were measured in the base pairs that had to be changed between the structure of the RNA occupying each cell and the target (base pair distance). From the eligible cells one was chosen randomly weighted by the relative fitness of the overgrowing cell. The relative fitness p(d_i)was determined with the following equation:

$$p(d_i) = \frac{e^{\frac{-d_i}{d}}}{\sum_{i=1}^{N} e^{\frac{d_i}{d}}}, \qquad \textit{(Equation 1)}$$

where d_i is the base pair distance of the structure from the target and d is the mean of this distance within the eligible cells. When a cell was overgrown by one of its neighbours the neighbour was copied with mutations, the mutation rate was 0.01 per base and the replacing base was chosen randomly with equal chance from all 4 canonical bases, so there was a 25% chance the base was replaced with itself. This setup was simulated for the target structure ".......((((....))))(((((........)))))............" (Figure 1).

There were five types of simulations: i) no adapters and no mixing, ii) no adapters and the grid is perfectly mixed every timestep, iii) every cell starts with the same adapter and no mix, iv) every cell starts with the same adapter and the grid is perfectly mixed every timestep, v) every cell starts with a random adapter and the grid is not mixed. Each of these started with a population of the same RNAs. 10 different initial sequences were used (Table 1).

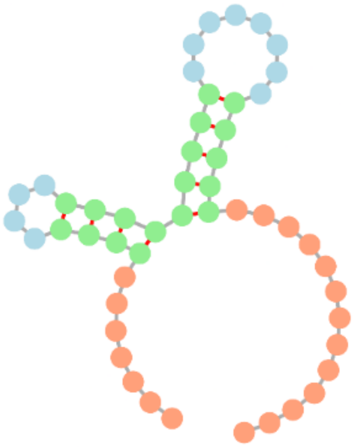
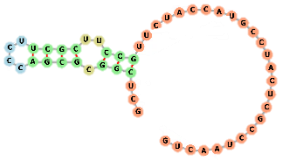
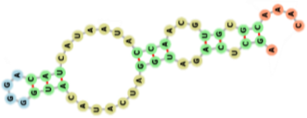
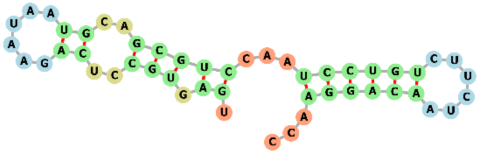
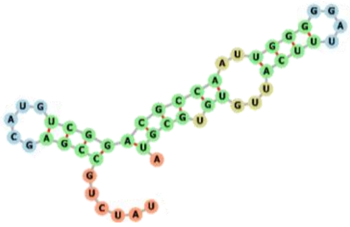
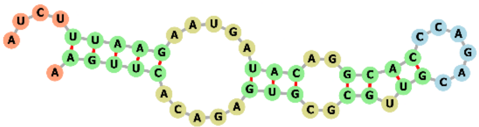
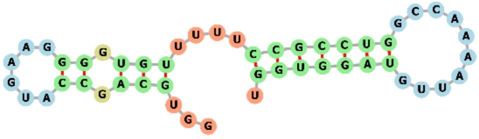
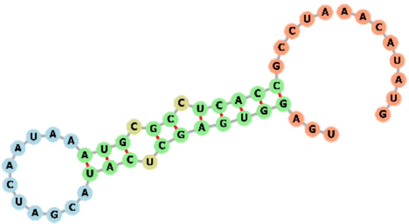
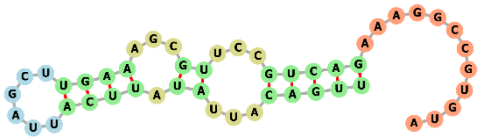
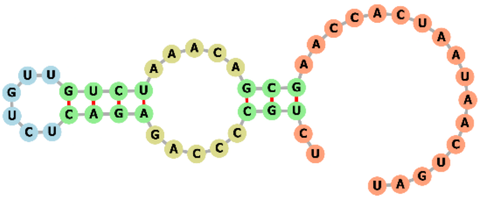
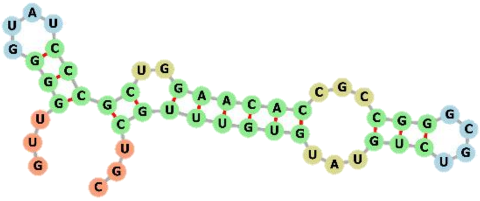


Figure 1: The target structure used for the simulations. The image was produced with the forna program of the ViennaRNA web services.

Sequence	Secondary structure
GCUCGGCGCGACCCUUCGCUUCCGUUCUACCAUGCCUACUCGCCUAACUG	
AGCUCAGAUGGAUCAUACAUGGGGACAUCAUAAUACCAACGUGCGCAAAC	
UGAGUGCCUCAGAAUAAUGCAGCGUCCAAUCCUGUCUUCUAACAGGAACC	
UAUCUGCCGAGCAUGUCGGACGCCAAUUGGGGGAUUUCAUUGUGUGCGUA	
AUCUUUAAGAAUGAUACAGGCACCCAGACGUUGCGCGUGAGACACUUGAA	
GGUGCAGCCAUGAAGGGGUGUUUUUCCGCCUGGCCAAAAUUGUAGGUGGU	
UGAGGUGAGCUCAUACGAUCAAUAAAUGCGCCUCACCGCCUAAACAUAUG	
UUGACAUUAUAUUCAUUAGCUUGAAAGCGUUCCGUCAGAAAGGCCGUGUA	
UCUGCCCCAGAGACUCUGUUGUCUAAACAGCGAACCACUAAUAACUGAUG	
GUUGGGGUAUCCCGCUGGAACACCGCCGGGCGUCUGUAUGUGUUUGCUGC	

Tabe 1: The RNAs used as initial population.

In the simulations that had adapters a few new rules were introduced. The folding with adapters was done as follows. First the adapter was folded. If it had some kind of hairpin structure, then the non-paired region of the hairpin was used as a binding site. Following this the large RNA was searched for the alignment with the adapter's binding site, that had the lowest, negative minimal free energy. The bases that were bound by the adapter were ignored during the folding of the large RNA, thus creating a potentially different structure to that without an adapter. All the folding were done with a free energy minimalizing algorithm that was developed for the project (see in the references).

A cell only used the adapter for folding, if the resulting structure was closer to the target, than the structure without an adapter. During the overgrowth phase the adapter also got copied and mutated. After, the cells chose a random neighbouring cell and tried its adapter if this adapter was better than the cells own, then it was copied exactly to the cell.

Each simulation was done 25 times for the 10 different starting structures.

Results

Comparison of the different simulation types for each different initial populations show that in the majority of the cases the simulations with adapters had a higher chance of finding the target (Figure 2). On Figure 2 it is obvious that when the population was allowed to use adapters, in the majority of cases if it found the target, also more than 10% of the population used adapters to fold into the target structure. One can also see that in 4 time out of the 10 both starting with the same adapter and starting with random adapters are better, than going without adapters. Not using the adapters was beneficial in only two cases.

On Figure 3 is the connection between the type of simulation and the mean time to find the target structure in generations. Overall the control was the slowes and using the same adapters in the start and mixing the grid every timestep proved to be the fastest. The other three scenarios took were fairly the same time on average and the were in the middle compared to the "extremities".

In order to quantify the effect of adapters on sequence variation the the mean Hamming-distance between the sequences that had the target structure was calculated for each starting sequences (Table 2). The simulations were the sub.population with the target structure used at most 10% of the adapters, this sequence variety was quite small. In all but two cases the simulations where the adapters were used in a significant portion of the population for folding into the target the sequence variety was higher than without any adapters. It should be noted that the 2 simulations where the control had higher variety are not the same where the control had a higher chance of finding the target (Compare Figure 2 to Table 2).

Sadly at the end of the project a bug was discovered in the implementation of Equation 1. This bug effected cells and at least one of their neighbours had the target structure in the overgrowth step. Essentially it caused a right bottom corner to left top corner directed migration of patches of cells with the target structure. This phenomenon could not effect the chane or time to find the target, but could effect the end simulation sequence variation. Therefore the non-mixed control and the non-mixed (i), same starting adapter (iii) simulations were run again, 5 times for each starting sequence. The sequence varieties for those simulations can be seen in Table 3. Here we can see that only with half of the sequences was the control worse than the populations where more than 10% of the sequences used an adapter to fold into the target structure. It is important to recognise that these samples are only the fifth of the original comparisons.

Summary

All in all there is some weak evidence that point towards how adapters can effect the evolution of functional RNA populations. They have the ability to increase the chance of finding a target structure. I think this is because they enable folding into tha taget structure for sequences that would not have that secondary structure otherwise. But it is crutial to note that these conclusions are based on a relatively small sample and with only one target structure.

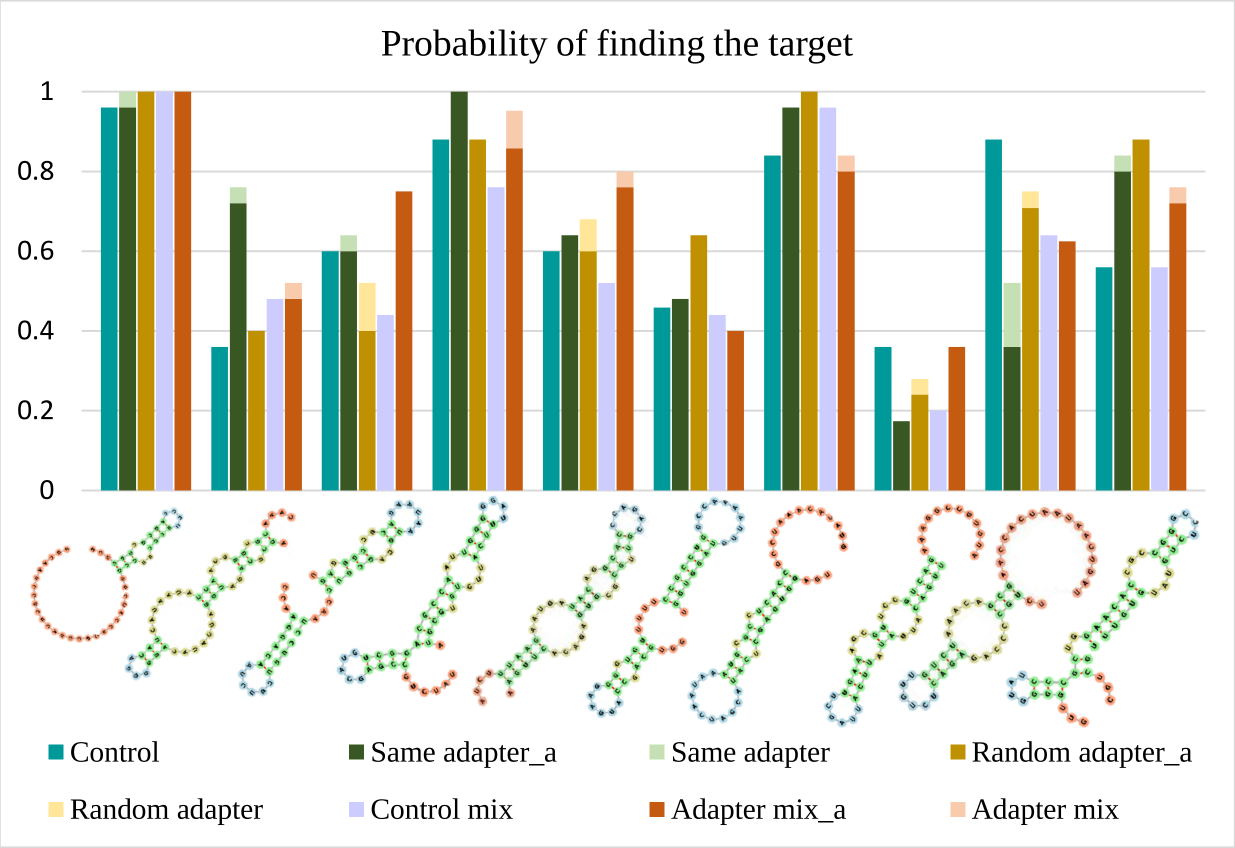


Figure 2: The probability of finding the target depending on the initial population and simulation type. On the bars belonging to the simulations with adapters we can see two colours. The darker colours represent the proportions of simulation where at the 100th timestep more than 10% of the RNAs with the target structure used adapters for folding (x_a). And with paler colour are the simulations where at most 10% of the target sequences were folded with an adapter.

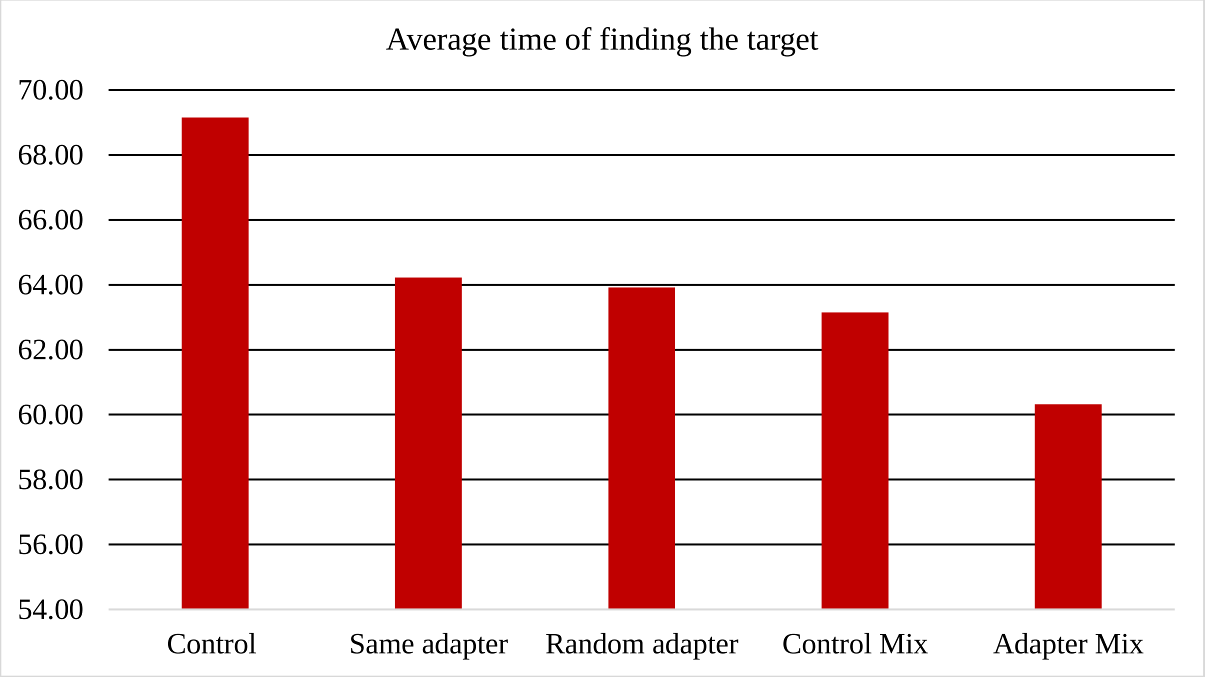


Figure 3: The mean time in generations to find the target structure in the different simulation types.

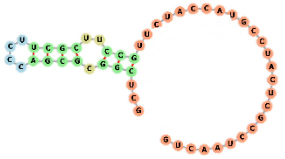
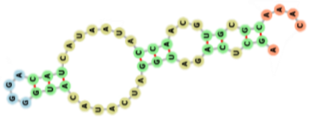
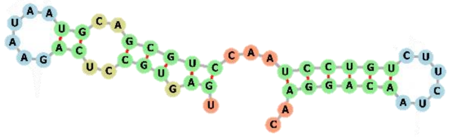
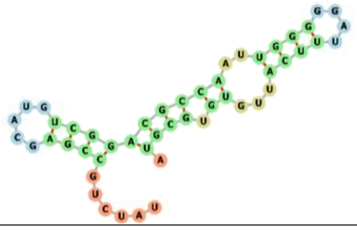
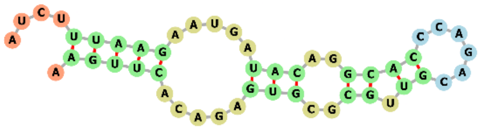
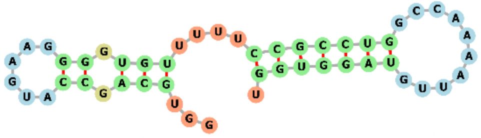
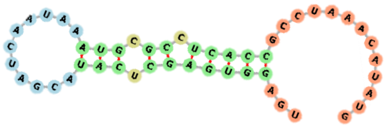
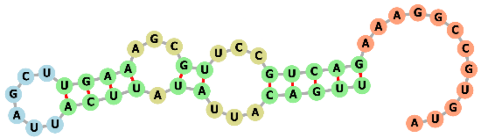
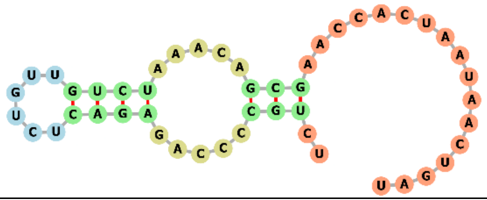
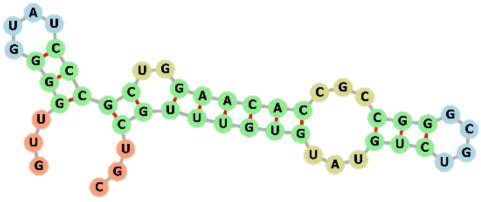
Initial RNA	Ctrl.	A > 10%	A <= 10%
	18.57	18.70	15.93
	11.52	10.68	1.01
	9.92	10.19	3.64
	8.47	13.61	NaN
	9.31	8.08	0.00
	6.98	7.90	NaN
	9.92	14.09	NaN
	8.16	11.18	2.21
	9.95	13.56	2.53
	8.35	11.17	0.00

Table 2: In this table are the sequence variations of two simulation types. In the first column are the initial sequences and their structures. In the "Ctrl" column the mean sequence variety at the last timestep of the simulation which did not use adapters and was not mixed (i). In the "A > 10%" and the "A <= 10%" columns are the mean sequence varieties for the simulation in which the same adapter was used as starting adapter and was not mixed (iii). The "A > 10%" column contains the simulations where at the end more than 10% of the sub-population with the target structure used its adapter for folding. In the "A <= 10%" column are the simulation were at most 10% of the sub-population with the target structure used its adapter for folding. "NaN" indicates that for that cell was no available data, because no simulations fell into that category.

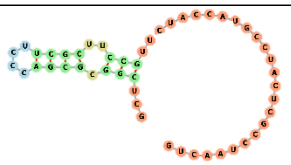
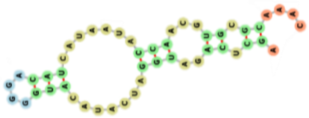
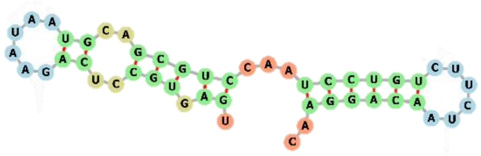
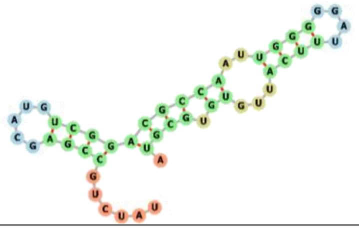
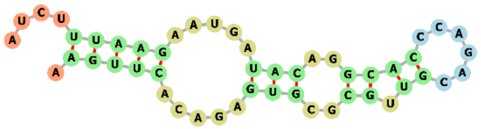
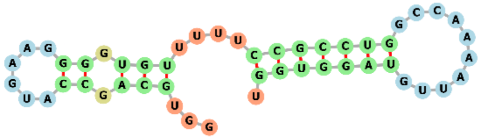
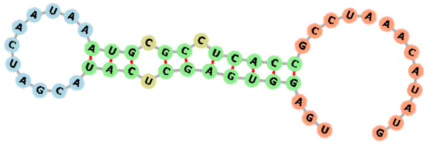
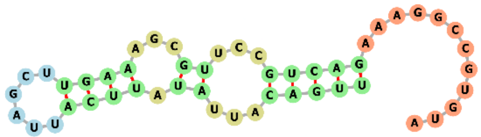
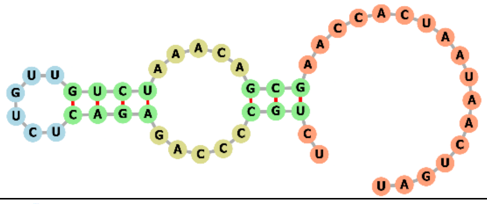
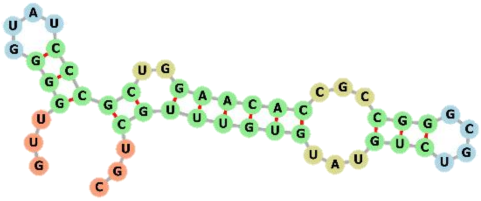
Initial RNA	Ctrl.	A > 10%	A <= 10%
	16.14	17.63	NaN
	6.94	9.97	NaN
	17.45	11.02	0.80
	8.63	12.06	0.69
	8.35	6.86	NaN
	9.20	7.35	NaN
	4.17	12.51	NaN
	10.05	0.84	2.91
	11.57	9.10	0.00
	8.95	9.46	NaN

Table 3: The columns are identical with Table 2. The sequence variety values for the simulations where the bug regarding the implementation of Eq. 1 was corrected.

References

de Boer, F. K., & Hogeweg, P. (2012). Less can be more: RNA-adapters may enhance coding capacity of Replicators. *PLoS ONE*, *7*(1). https://doi.org/10.1371/journal.pone.0029952

Gilbert, W. (1986). Origin of life: The RNA world. *Nature*, *319*(6055), 618–618. https://doi.org/10.1038/319618a0

Kun, Á., Szilágyi, A., Könnyű, B., Boza, G., Zachar, I., & Szathmáry, E. (2015). The dynamics of the RNA world: Insights and challenges. *Annals of the New York Academy of Sciences*, *1341*(1), 75–95. https://doi.org/10.1111/nyas.12700

Kóródi, L. (2024, January 24). *IAM-Locy/RNAfolding: JavaScript module for RNA secondary structure prediction*. RNAfolding. https://github.com/Iam-Locy/RNAfolding

van Dijk, B. (2022, February 6). *Bramvandijk88/cacatoo: A javascript library for building, exploring, and sharing spatially structured models of Biological Systems*. Cacatoo. https://github.com/bramvandijk88/cacatoo