# The necessary emergence of structural complexity in self-replicating RNA populations

Carlos G. Oliver[1,*], Vladimir Reinharz[2,*], Jérôme Waldispühl[1,†]

[1] School of Computer Science, McGill University, Montreal, Canada

[2] Department of Computer Science, Ben-Gurion University, Beer-Sheva, Israel

November 13, 2017

**Abstract**

The RNA world hypothesis relies on the ability of ribonucleic acids to replicate and spontaneously acquire complex structures capable of supporting essential biological functions. Multiple sophisticated evolutionary models have been proposed, but they often assume specific conditions. In this work we explore a simple and parsimonious scenario describing the emergence of complex molecular structures at the early stages of life. We show that at specific GC-content regimes, an undirected replication model is sufficient to explain the apparition of multi-branched RNA secondary structures – a structural signature of many essential ribozymes. We ran a large scale computational study to map energetically stable structures on complete mutational networks of 50-nucleotide-long RNA sequences. Our results reveal regions of the sequence landscape enriched with multi-branched structures bearing strong similarities to those observed in databases. A random replication mechanism preserving a 50% GC-content suffices to explain a natural drift of RNA populations toward complex stable structures .

*These authors contributed equally to the work.

†To whom correspondence should be addressed. Tel: +1 514-398-5018; Email: `jerome.waldispuhl@mcgill.ca`

# 1 Introduction

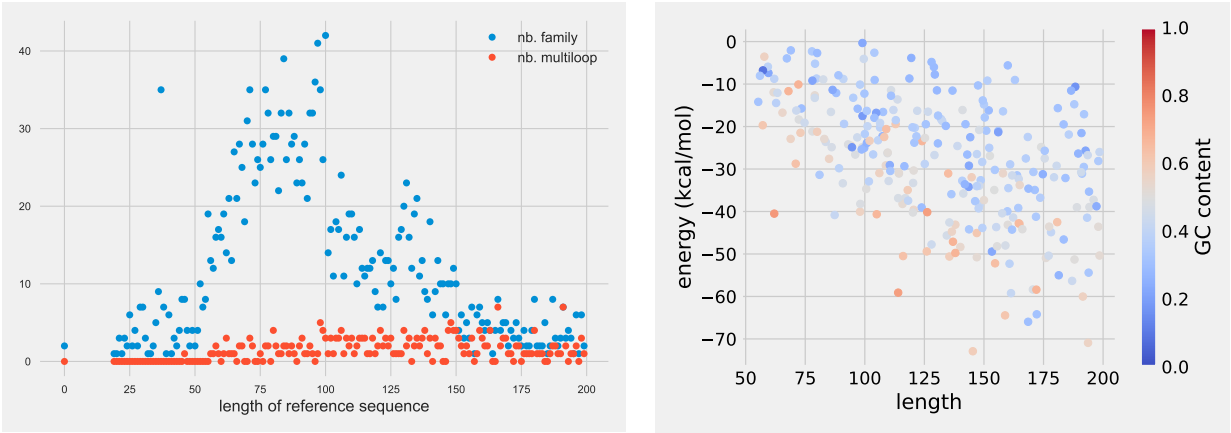
RNA are versatile molecules that can fulfil virtually all fundamental needs and functions of the living, from storing information to catalyzing chemical reactions and regulating gene expression. The RNA world hypothesis[1,2] builds upon this observation to describe a scenario of the emergence of life based on RNAs. Recent experimental studies showing the remarkable adaptability of nucleic acids contributed to strengthen this hypothesis[3,4] and thus revived the interest of the scientific community for this theory.

Non-coding RNAs acquire functions through complex structures. A theory describing how nucleic acids evolve to "discover" these functional structures is thus a milestone toward a validation of the RNA world hypothesis[5].

Since the first analysis of RNA neutral networks[6], computer simulations are the method of choice for characterizing the evolutionary landscape and population dynamics of RNA molecules. This situation is motivated by three factors. First, RNA secondary structures are more conserved than sequences and provide a reliable signature of RNA function[7]. Next, these secondary structures can be reliably predicted from sequence data only[8], allowing fast and accurate prediction of phenotypes (i.e. secondary structure) from genotypes (i.e. sequence). Finally, the computational power available for computational simulations is growing increasingly faster, so is the breadth and depth of the studies.

A large body of literature has been dedicated to the computational analysis of RNA sequence-structure maps (i.e. genotype-phenotype maps) and properties of RNA populations evolving under natural selection. In a seminal series of papers P. Schuster and co-workers set up the basis of a theoretical framework to study evolutionary landscape of RNA molecules, and used it to reveal intricate properties of networks of sequences with the same structure (a.k.a. neutral networks)[6,9–13]. This work inspired numerous computational studies that refined our understanding of neutral models[14–18], as well as kinetics of populations of evolving nucleic acids[19–21].

Many essential molecular functions are supported by nucleic acids with complex shapes (e.g. 5s rRNA, tRNA, hammerhead ribozyme). Here, we define a complex structure as a secondary structure that contains a multi-loop. This feature is relatively common even for short RNAs. Indeed, an analysis of the consensus secondary structures available in the Rfam database reveals that multi-




(a) (b)

Figure 1: Statistics on Rfam families[7]. On the left, the graph plots the number of families with respect to the average length of the sequences in these families. Red dots show the numbers of families with a consensus structure that contains a multi-loop, while blue dots show those without. On the right, we plot the average folding energy and length of sequences for each Rfam families having multi-branched consensus structures. The color indicates the average GC content of the family.

loops can be found in approximately 10% of RNA families whose average size of sequences ranges from 50 to 100 nucleotides (See **Fig. 1a**).

These observations contrast with earlier studies by S. Manrubia and co-workers that revealed that the vast majority of predicted minimum free energy (MFE) secondary structures obtained from an uniform sampling of RNA sequences are stem-like structures (i.e. no multi-loop)[22]. Although the size of sequences analyzed in the latter study was limited to 35 nucleotides, the sparsity of multi-branched structures in the sequence landscape calls for a clarification of the evolutionary mechanisms that enabled the emergence of structural complexity in a prebiotic world.

Previous computational studies that investigated the properties of RNA sequence-structure maps showed that neutral networks percolate the whole sequence landscape[23,24]. Even though this property undeniably augments the accessibility of target structures, the size and connectivity of neutral networks also decreases drastically with the complexity of the structure.

In the most commonly accepted scenarios, the establishment of a stable, autonomous, and functional self-reproductive molecular system subject to natural selection, relies on the presence of polymerases[5]. Such molecules are long (200 nt.) and thus unlikely to be discovered randomly. Instead, it has been suggested that evolution proceeded by stages[25]. Polymerases were assembled

from smaller monomers (∼50 nt.) that are more likely to emerge from prebiotic chemistry[26,27]. At this point, and not before, parallel and independent natural selection of individual molecular structures could be triggered.

Interestingly, *in vitro* experiments revealed the extreme versatility of random nucleic acids[28–30]. Other studies have also suggested that essential RNA molecules such as the hammerhead ribozyme have multiple origins[31]. All together, these observations reinforce the plausibility of a spontaneous emergence of multiple functional sub-units. But they also question us about the likelihood of such events and the existence of intrinsic forces promoting these phenomena.

Various theoretical models have been proposed to highlight mechanisms that may have favoured the birth and growth of structural complexity from replications of small monomers. Computational studies have been of tremendous help to validate these theories and quantify their impact. In particular, numerical simulations enabled us to explore the effects of polymerization on mineral surfaces[32,33] or the importance of spatial distribution[34]. Another important aspect of early life models is the tradeoff between stability and structural complexity. Stable folds often lack the complexity necessary to support novel functions but are more resilient to harsh pre-cellular environments [35]. Still, the debate about the necessity for such hypotheses remains open.

In this work, we show that structural complexity can naturally emerge without the help of any sophisticated molecular mechanisms. We reveal subtle topological features of RNA mutational networks that helped to promote the discovery of functional RNAs at the early stages of the RNA world hypothesis. We demonstrate that in the absence of selective pressure, self-replicating RNA populations naturally drift toward regions of the sequence landscape enriched in complex structures, allowing for the simultaneous discovery of all molecular components needed to form a complete functional system.

For the first time, we apply customized algorithms to map secondary structures on all mutant sequences with 50 nucleotides[36,37]. This approach considerably expands the scope and significance of comprehensive RNA evolutionary studies that were previously limited to sequences with less than 20 nucleotides[11,38], or restricted to explore a small fraction of the sequence landscape of sequences[18,22]. This technical breakthrough is essential to observe the formation of complex multi-branched structures often used to carry essential molecular functions that cannot be assembled on shorter sequences.

Our simulations reveal the unexpected presence of a large pool of remarkably stable multi-branched structures in a region of the RNA mutational landscape characterized by an average distance of 30 to 40 mutations from a random sequence, and a balanced GC content (i.e. 0.5). Strikingly, these multi-branched RNAs have similar energies ($-15 \pm 5\,\mathrm{kcal\,mol^{-1}}$) to those observed in the Rfam database[7] on the same length scale (See **Fig. 1b**).

We compare these data to populations that evolved under a selective pressure eliciting stable structures. Although this evolutionary mechanism shows a remarkable capacity to quickly improve the stability of structures, it fails to reproduce the structural complexity observed in RNA families of similar lengths.

Finally, we show that a population of RNA molecules replicating itself randomly with random errors but preserving a balanced GC content[39], naturally evolves toward regions of the landscape enriched with multi-branched structures potentially capable of supporting essential biochemical functions. Our results argue for a simple scenario of the origin of life in which an initial pool of nucleic acids would irresistibly evolve to promote a spontaneous and simultaneous discovery of the basic bricks of life.

# 2 Results

## 2.1 Our approach

We apply two complementary mutation space search techniques to characterize the influence of sampling process to the repertoire of shapes accessible from an initial pool of random sequences (See **Fig. 2**). Importantly, our analysis explicitly models the impact of the GC content bias. Our first algorithm `RNAmutants` enumerates all mutated sequences and samples the ones with the *globally* lowest folding energy[36]. It enables us to calculate the structures accessible from a random replication process. In contrast, our other algorithm named `mateRNAl`, has been developed for this study to simulate the evolution of a population of RNA sequences that preferentially selects the most stable structures under nucleotide bias.

Using `RNAmutants`, we compute for the very first time, the energy landscape of the complete mutational network of RNA sequences of length 50 preserving the GC content (See **Fig. 3a**). This result contrasts with previous studies that used brute-force approaches to calculate the complete

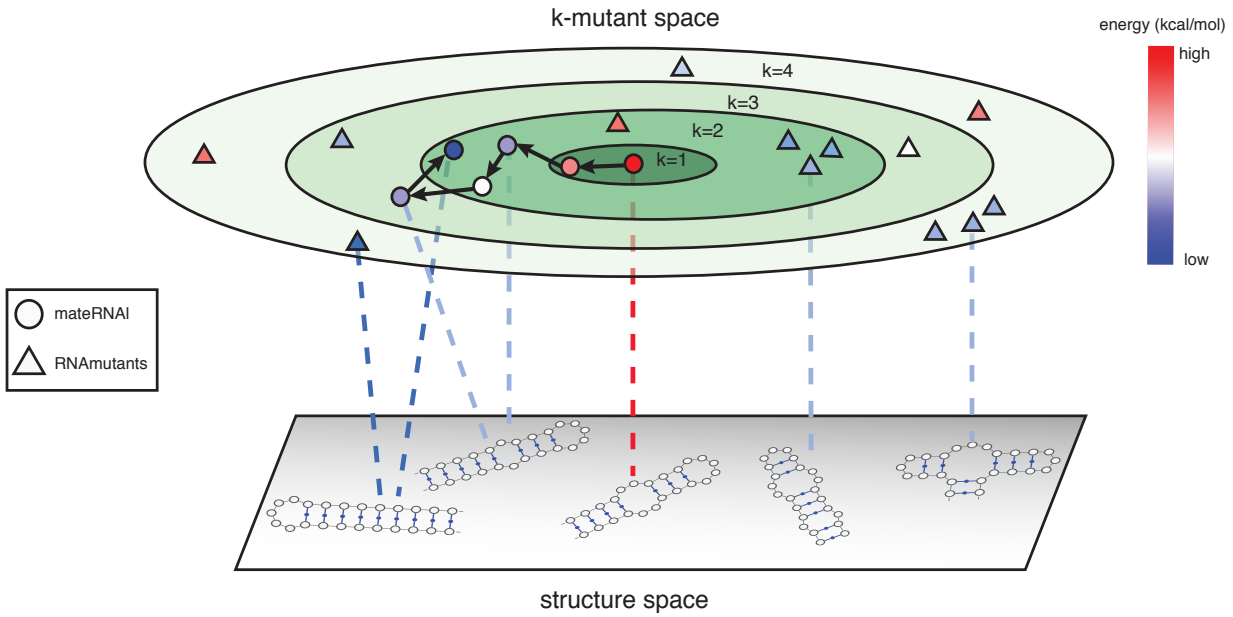



Figure 2: *Illustration of mutational sampling methods.* Concentric ellipses represent the space of sequence k-mutant neighbourhoods around a root sequence pictured at the centre of the ellipses. Each ring holds all sequences that are $k$ mutations from the root. We show the contrast between an evolutionary trajectory (`mateRNAl`) along this space and mutational ensemble sampling (`RNAmutants`) represented respectively as circles and triangles. The layer below the mutational space represents the space of all possible secondary structures, and dotted lines illustrate the mapping from sequences to structures. The colour of the sampled mutants denotes the energy of the sequence-structure pair sampled. In both sampling methods, sequence-structure pairs with lower energies are favoured. Evolutionary sampling is always limited to explore sequences accessible from the parental sequence and so we have arrows pointing from parents to children over various generations yielding an adaptive trajectory. `RNAmutants` considers the entire ensemble of $k$ mutants to generate independent samples of stable sequence-structure pairs and thus reveals features such as complex structures that are hard to reach by local methods such as `mateRNAl`.

sequence landscape, and were therefore limited to sizes of 20 nucleotides. This technical breakthrough is essential because multi-branched structures occur only on RNAs with sizes above 40 (See **Fig. 1**)

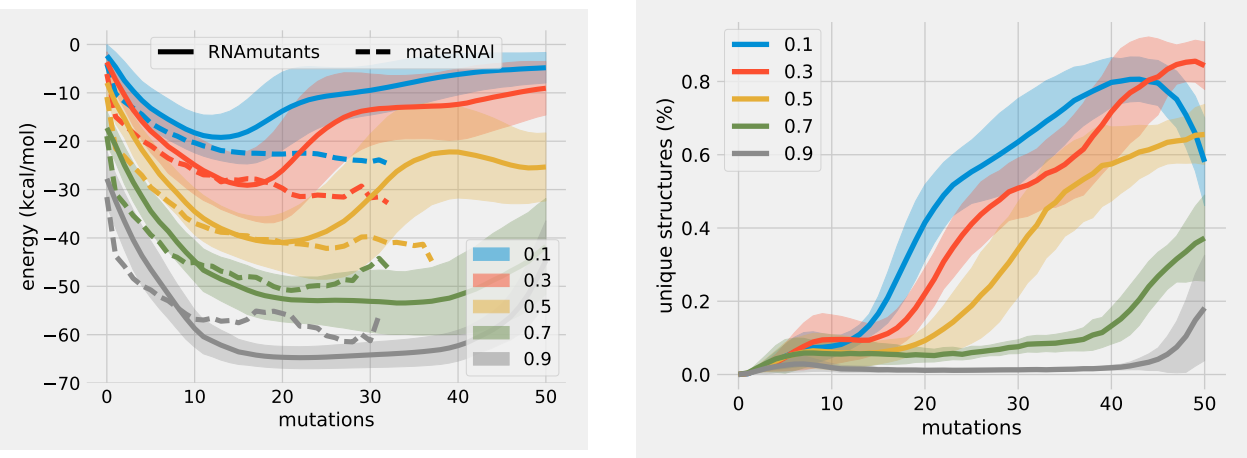
More precisely, given an initial sequence (i.e. the *seed*), `RNAmutants` calculates mutated sequences with the lowest folding energies among all possible secondary structures (i.e. no target secondary structure). Importantly, `RNAmutants` allows us to control the GC content of mutated sequences[37] (fraction of bases in a sequence that are either G or C). We ran `RNAmutants` on 20 independently generated random sequences for each GC content regime in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ ($\pm 0.1$). At each mutational distance (i.e. number of mutations from the seed) from 0 to 50, we sample from the Boltzmann distribution approximately 10 000 sequence-structure pairs with the target GC content. It is worth mentioning that the secondary structures sampled by `RNAmutants` are not necessarily MFE structures of the sequences associated with. However, our data shows that the vast majority of these sampled structures are very close to their MFE counterpart (See **Supp. Fig. S1**). This observation enables us to simplify analysis by retaining MFE structures and produce data that is directly comparable with `mateRNAl`.

## 2.2 Energy landscape of RNA mutational networks

We start by characterizing the distribution of folding energies of stable structures accessible from random seeds in the mutational landscape using `RNAmutants`. Our simulations show that, initially, increasing mutational distances result in more stable structures at all GC content regimes (See **Fig. 3a solid line**). In particular, we observe that the folding energies of the samples represent at least 80% of the global minimum energy attainable over all $k$ mutations for all GC contents within less than 10 mutations from the seed (see **Supp. Fig. S2**). This finding suggests that over short evolutionary periods, mutations are likely to play an important stabilizing role.

By contrast, at larger mutational distances (i.e. 15 mutations and above) we notice an increase in the ensemble energy for all GC content regimes. This behaviour is likely due to the exponential growth in sequence space that accompanies higher mutational distances, which results in a more uniform sampling of the ensemble. In this case, lower energy sequences with high Boltzmann weights become less represented among more abundant and less stable sequences (See **Fig. S1**).

As expected, we observe that the nucleotide content is an important factor in determining the




(a) (b)

Figure 3: (a) Energy of `RNAmutants` and `mateRNAl` mutational landscape. Shaded regions include one standard deviation of `RNAmutants` energy per mutational distance. Dashed lines mark mean values for `mateRNAl` energies binned by mutations from starting sequence using mutation rate $\mu = 0.02$. (b) Fraction of unique structures at every $k$ neighbourhood found by `RNAmutants`.

stability of achieved sequence-structure pairs across mutational networks. The accessible sampled energies are strongly constrained by the allowed GC content of sequences in all GC content regimes, whereby higher GC contents favor the sampling of more stable states. However, despite these constraints, all mutational ensembles effectively produce more stable states than the initial state.

Having established that mutational neighbourhoods of random sequences yield stable and low energy states, it is important to also understand the structural features of these states. This is of particular interest trying to provide an account of the emergence of functional and complex RNAs. While previous studies on the structure of short randomly sampled RNA sequences have shown that simple hairpin structures dominate the landscape, we find that for longer molecules the landscape of stable mutant neighbourhood reveals ensembles that are well populated with diverse structures (See **Fig. 3b**). Interestingly, for all GC content sampling and particularly at low to intermediate GC contents, we find that after an initial stabilization period at short mutational distances ($\sim 10$ mutations) more complex structures begin to emerge. This sudden and unexpected change of regime seems to correlate with the decrease of the average folding energies observed in **Fig. 3a**.

We can observe this change as an increase in the number of structures with internal loops (**Fig. S5**) and remarkably, multi-loops (**Fig. 4a**). This finding is in good qualitative agreement

with databases of evolved structures which contain structures with more diverse motifs than has been previously estimated from pools of random RNAs.

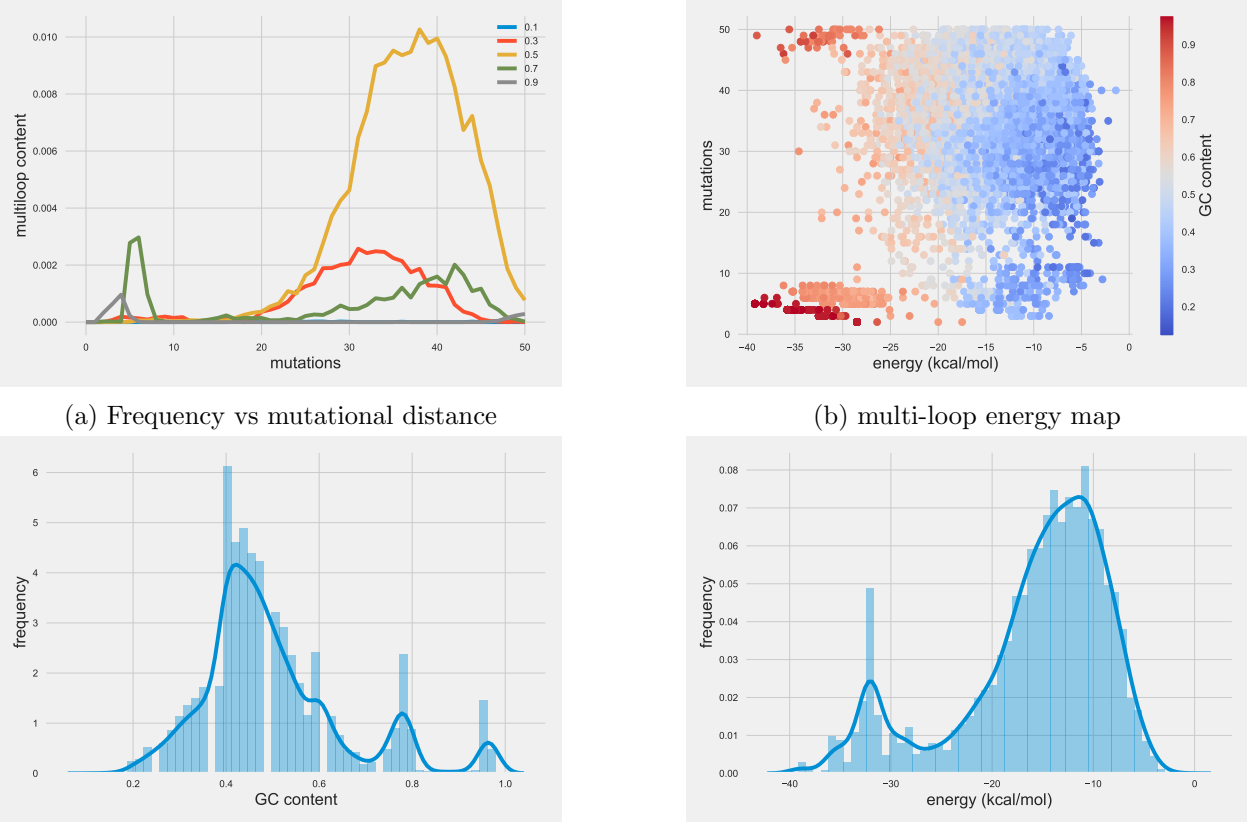
These secondary structure features (i.e. internal and multi-loops) are key components of RNAs that allow for higher order interactions that can support catalytic function. Although multi-loops occur at relatively low frequencies in the `RNAmutants` mutational landscapes (at most 0.01), we focus on them for their high degree of structural complexity and importance to many functional RNAs observed in vivo such as the hammerhead ribozyme.

Across all runs, we sampled a total of 9419 sequences containing a multi-loop (no more than 1 multi-loop per structure was ever observed as is to be expected for such length scales). Interestingly, we find that unlike internal loops, multi-loops occur under very specific conditions in our sampling. We identify a clear surge of multi-loops frequency at distances at a mutational distance of $\sim 35$ (See **Fig. 4a**), with a mean GC content of $\sim 0.45$ (See **Fig. 4c**). Furthermore, their energy distribution is tightly centered around $\sim -15\,\mathrm{kcal\,mol^{-1}}$ (See **Fig. 4d**). These values are remarkably close to the those of multi-branched structures of similar lengths observed in the Rfam database (See **Fig. 1b**). In particular, the latter shows a clear bias toward medium GC contents as we identified 148 Rfam families with multi-loops with a GC content of 0.5 (among all Rfam families with sequences having at most 200 nucleotides), but only 80 with a GC content 0.3 and 40 with a GC content of 0.7. This serves as further evidence that GC content is an important determinant of the evolution of structural complexity. It also appears that these features of multi-loops are a general property of the distribution of multi-loops in the mutational landscape as the entropy of the set of sequences containing multi-loops is 0.945.

Eventually, we also note a smaller peak of multi-loop occurences closer from the seed sequences ($\sim 6$ mutations) for higher GC contents around 0.7. Interestingly, with folding energies ranging from $-25$ to $-40\,\mathrm{kcal\,mol^{-1}}$, these multi-branched structures are significantly more stable than those present in the main peak (See **Fig. 4b**). This is also in agreement with the energies observed in the Rfam database for structures within this range of GC content values (See **Fig. 1b**).

### 2.3 An energy-based evolutionary model

Our `RNAmutants` simulations suggest that mutational networks traversed in an energy based manner can yield stable and diverse structures. We revealed that multi-branched structures reside in ~~specific~~




(a) Frequency vs mutational distance

(b) multi-loop energy map

(c) GC content distribution

(d) Energy distribution

Figure 4: Analysis of the distribution of multi-branched structures in the RNA mutational landscape. (a) The frequency of multi-branched structures with respect to the number of mutations from the seed sequence. (b) Plot of folding energies and GC contents of each individual multi-branched structure. (c) Distribution of the GC content of multi-branched structures. (d) Distribution of folding energies of multi-branched structures.

regions of the mutational landscape at fixed mutational distances (primarily at 30–40 mutations from random seeds) and GC contents (0.5 for the majority of structures). At this point, our main question is to determine if a natural selection process, independent of a particular target, is capable of reaching these regions.

To address this question we build an evolutionary algorithm named `mateRNA1`, where the fitness is proportional to the folding energies of the molecules. Intuitively, `mateRNA1` enables us to simulate the behaviour of a population of RNAs selecting the most functional sequences (i.e. most stable) regardless of the structures used to carry the functions. These selected structures are therefore by-products of intrinsic adaptive forces.

We start all simulations from random populations of size 1000 and sequences of length 50, and

performed 50 independent simulations for each GC content and various mutation rates. All simulations were run for 1000 generations. **Fig. 3a** shows the mean of the folding energies of `mateRNAl` sequences binned by their distance (i.e. number of mutations) from the initial population. We find that populations of random RNA sequences are able to quickly find low energy solutions (less than 100 generations and $\sim 12$ mutations from the initial populations; See **Fig. S4a**). Furthermore, we note that although `mateRNAl` selects on average more stable sequences in the vicinity of the seeds (i.e. the initial population) than `RNAmutants`, this does not hold for higher mutational distance and higher GC contents.

More precisely, we observe that at short mutational distances, evolutionary approaches are able to harness the larger initial variation present (initial population size of 1000 versus a single starting seed sequence) to rapidly identify low energy structures. However, with greater mutational distances, and as selection sorts through the population, diversity is depleted and populations lack the necessary variation to overcome energy barriers and obtain energies reached by RNAmutants. This behaviour is also reflected in the limited mutational depth accessible by maternal (See **Fig. 3a** and **Fig. 5**).

Interestingly, we find that varying the mutation rate does not have a strong effect on the energy of populations obtained (See **Fig. S4b**). We see that the average population energy remains within $\pm 5\,\mathrm{kcal\,mol^{-1}}$ for all mutation rates except the highest of 0.1. This apparent property of energy based evolutionary models suggests that the search for stable structures is flexible to external conditions such as varying mutation rates and may provide a mechanism for better exploring phenotype space without sacrificing stability.

While we observe very efficient energy optimization from the natural selection process, `mateRNAl` fails to generate the structural complexity found in databases , but successfully uncovered by `RNAmutants` (See **Table. S1**). This is likely due to the rapid depletion of diversity inherent to selection under fixed population sizes and the strong selection for highly stable yet simple folds. However, our model does allow for some control over the degree of complexity obtained. Notably, we see that the mutation rate has a strong impact on the mean stack content, internal loop, and multi-loop content across populations whereby higher mutation rates promote the discovery of more complex structures (See **Fig. S6**). This increase in occurrence of complex motifs with high mutation rates, notably multi-loops, does not result in the fixation of these structures in the

population, rather in their occasional sampling and subsequent disappearance.

### 2.4 An undirected evolutionary scenario

We are now showing that random replication *without natural selection* is sufficient to explain the diversity of structures observed in RNA databases, and further the emergence of RNA structural complexity. We consider a simple model in which RNA molecules are duplicated with a small error rate, but preserving the GC content[39]. In our simulations, we use an error rate of 0.02 to allow a immediate comparison of the number of elapsed generations. We also apply identical transitions and transversions rates. Under these assumptions, we can directly compute the expected number of mutations in sequences at the $i^{th}$ generation (see Methods). **Fig. 5** shows the results of this calculation for GC content biases varying from 0.1 to 0.9. Strikingly, our data reveals that after a short initialization phase (i.e. after $\sim 50$ generations), sequences with a GC content of 0.5 have on average slightly more than 35 mutations. This observation is in perfect adequacy with the peak of multi-branched structures identified in **Fig. 4a** and **Fig. 4c**. It follows that a simple undirected replication mechanism, not subject to natural selection, is sufficient to explain a drift of populations of RNA molecules toward regions of the sequence landscape enriched with multi-branched structures.

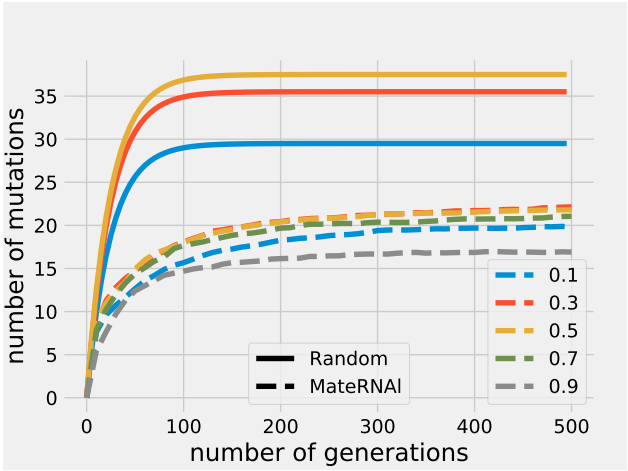



Figure 5: Number of mutations accumulated in populations evolving under a random replication model ("Random") and selection pressure ("`mateRNAl`")

The abundance of complex structures at large mutational distances (i.e. 30 to 40 mutations from the seed) could eventually be linked to larger sizes of the hamming neighbourhoods (See **Fig. S7**).

Nonetheless, here we also show that this phenomenon indeed coincides with an enrichment of stable multi-branched structures at specific GC contents (i.e. between 0.3 and 0.5).

In **Fig. 6**, we compare the average minimum free energy (MFE), as well as the frequency of multi-loops in MFE structures, of sequences sampled from a uniform distribution of mutations ("Random") to sequences sampled from the low-energy ensemble ("RNAmutants").

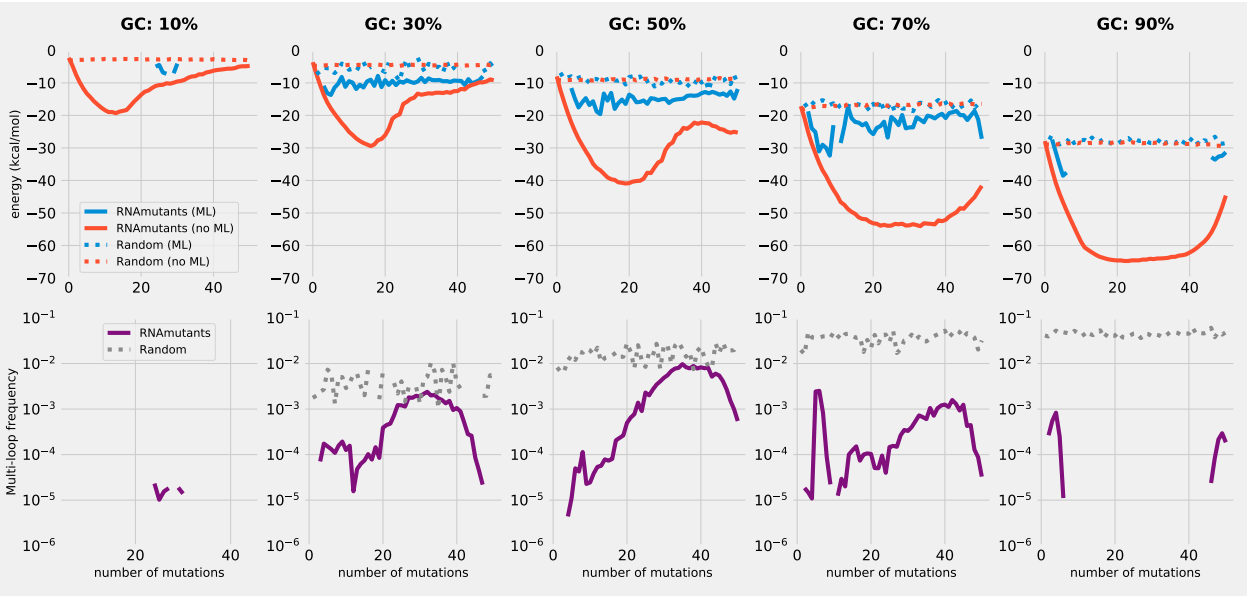



Figure 6: Analysis of multi-loop distributions in uniform and low energy populations. First row: average minimum free energies of uniformly sampled mutants ("Random"; dotted lines) and sequences from the low energy ensemble ("RNAmutants"; plain lines). Sequences with a minimum free energy structure having a multi-loop ("ML"; blue) are separated from the others ("no ML"; orange). Second row: frequency of multi-branched structures in the minimum free energy structure of uniformly sampled mutants ("Random"; plain grey lines) and and sequences sampled from the low energy ensemble ("RNAmutants"; dotted purple lines)

We uniformly sampled random sequences in each mutational neighbourhood and calculated their minimum free energy (MFE) secondary structures. Our data confirms previous observations made on shorter sequences showing that multi-branched structures are rare and relatively unstable in random populations, but that their frequency increases with higher GC content biases [22]. Strikingly, we note that the frequency of stable multi-branched structures in the low energy ensemble (i.e. "RNAmutants") is almost matching those obtained with random sequences at GC contents between 0.3 and 0.5 and mutational distances from 30 and 40 (second row in the third and fourth columns of **Fig. 6**), but with higher absolute values for higher GC content regimes (i.e. 0.5).

The analysis of folding energies shed a different light on this phenomenon. While the average

energies of multi-branched structures remains steady at all mutational distances, this is not the case of other structures with simpler architectures (first row of **Fig. 6**). Lower GC content regimes from 0.3 to 0.5 are characterized by a clear increase of average energies at mutational distances over 20 (i.e. MFE structures are less stable), which is not observed at higher GC contents. We conclude from these observations that the relative weight of multi-branched structures in the low energy ensemble increases due to a better (collective) resilience of this architecture to point-wise mutations and/or a more uniform distribution in the sequence landscape. In turn, it increases their density in the large/distant mutational neighbourhoods.

Eventually, we also distinguish a secondary peak of occurrences of multi-branched structures in the vicinity of the seeds (i.e. 5-10 mutations) at higher GC regimes (0.7). By contrast, this higher density appears to result from mutants folding with marginally lower energies. It suggests the presence of mutants with improved fitness to the structures of the seeds rather than a global enrichment of multi-branched structures in these neighbourhoods.

In conclusion, assuming that functional structures are preferentially fixed on stable structures available in the sequence landscape, our simulations suggests that GC content regimes of 0.5 favor the discovery of multi-branched structures. The probability of spontaneous emergence of complex structures in a random replication model may thus be higher than currently estimated.

## 3 Discussion

We provided evidence that in the absence of selective pressure the structure of the mutational landscape could have helped to promote the emergence of an RNA-based form of life. To support our hypothesis, we built a comprehensive representation of the mutational landscape of RNA molecules, and investigated scenarios based on distinct hypotheses.

Our results offer solid foundations to parsimonious evolutionary scenarios based on undirected molecular self-replications with occasional mutations. In these simple models, the GC content appears as a key feature to determine the probability of discovering stable multi-branched secondary structures. In particular, intermediate GC contents (i.e. 0.5) result in a drift of randomly replicating populations toward a sub-space of the evolutionary landscape uncovered by `RNAmutants` that drastically increases the probability of discovering thermodynamically stable complex shapes essential

for the emergence of life at the molecular level.

The preservation of intermediate GC content values appeared to us as a reasonable assumption, which could reflect the availability of various nucleotides in the prebiotic milieu. This nucleotide composition bias can be interpreted as an intrinsic force that favoured the emergence of life. It also offers novel insights into fundamental properties of the genetic alphabet[40].

In addition to this central reservoir of complex structures, our data also revealed the occasional presence of stable multi-branched structures in the vicinity of random sequences at GC content regimes of 0.7 (See **Fig. 4a** and **Fig. 6**). This finding is in agreement with previous theoretical studies that showed that neutral networks percolate the whole sequence landscape[23,24]. However, our simulations also suggest that a random replication model would only transiently occupy these regions, hence with significantly lower probabilities to find these structures. Eventually these close multi-branched structures appear to have a GC content ($\geq 0.7$) similar to those of RNA families with longer structures of more than 70 nucleotides (See **Fig. 1b**). In these particular cases, we conjecture that the structures were selected later in evolution by natural selection processes.

Eventually, our results could be used to put in perspective earlier findings suggesting that natural selection is not required to explain pattern composition in rRNAs[41]. Incidentally, these observations suggest further investigations into the role of more complex nucleotide distributions[42].

Our analysis completes recent studies that aimed to characterize fundamental properties of genotype-phenotype maps[43,44], and showed that their structure may contribute to the emergence of functional molecules[18]. It also emphasizes the relevance of theoretical models based on a thermodynamical view of prebiotic evolution[45].

The size of the RNA sequences considered in this study has been fixed at 50 nucleotides. This length appears to be the current upper limit for non-enzymatic synthesis[46], and therefore maximizes the expressivity of our evolutionary scenario. Variations of the sizes of populations or lengths of RNA sequences resulting from indels could be eventually considered with the implementation of dedicated algorithms[47]. Although we do not expect any major impact on our conclusions.

The error rates considered in this study were chosen to match values used in previous related works (e.g[48]). This choice is also corroborated by recent experiments suggesting that early life scenarios could sustain high error rates[49]. Nevertheless, lower mutation rates would only increase the number of generations needed to reach the asymptotic behaviour (See **Fig. 5**), and thus would

not affect our results.

Finally, we emphasize that our results do not preclude the existence of more advanced evolutionary mechanisms[32–34,50]. Nonetheless, they provide additional evidence supporting an RNA-based scenario for the origin of life, and can serve as a solid basis for further investigations of more sophisticated models.

# 4 Materials and Methods

## 4.1 Evolutionary Algorithm (`mateRNAl`)

### 4.1.1 Initial Population

Here we describe an evolutionary algorithm (EA) for energy-based selection with GC content bias. The algorithm is implemented in Python and is freely available at `http://jwgitlab.cs.mcgill.ca/cgoliver/maternal.git`.

We first define a population at a generation $t$ as a set $P_t$ of sequence-structure pairs. We denote a sequence-structure pair as $(\omega, s)$ such that $s$ is the minimum free energy structure on sequence $\omega$ as computed by the software `RNAfold` version 2.1.9[8]. Each sequence is formed as a string from the alphabet $\mathbb{B} := \{A, U, C, G\}$. For all experiments we work with constant population size of $|P_t| = 1000 \quad \forall \quad t$, and constant sequence length $\texttt{len}(s_i) = 50 \quad \forall s_i \in P_t$ . We then apply principles of natural selection under Wright-Fisher sampling to iteratively obtain $P_{t+1}$ from $P_t$ for the desired number of generations in the simulation.

Sequences in the initial population, i.e. generation $t = 0$, are generated by sampling sequences of the appropriate length uniformly at random from the alphabet $\mathbb{B}$.

### 4.1.2 Fitness Function

In order to obtain subsequent generations, we iterate through $P_t$ and sample 1000 sequences with replacement according to their relative fitness in the population. Selected sequences generate one offspring that is added to the next generation's population $P_{t+1}$. Because we are sampling with replacement, higher fitness sequences on average contribute more offspring than lower fitness sequences. The relative fitness, or reproduction probability of a sequence $\omega$ is defined as the

probability $F(\omega, s)$ that $\omega$ will undergo replication and contribute one offspring to generation $t+1$. In previous studies, $F(\omega_i, s_i)$ has been typically defined as a function of the base pair distance between the MFE structure of $\omega$ and a given target structure. However, in our model, this function is proportional to the free energy of the sequence-structure pair, $E(\omega, s)$ as computed by `RNAfold`.

$$F(\omega, s) = N^{-1} e^{\frac{-\beta E(\omega, s)}{RT}} \tag{1}$$

The exponential term is also known as the Boltzmann weight of a sequence-structure pair. $N$ is a normalization factor obtained by summing over all other sequence-structure pairs in the population as $N = \sum_{\omega', s' \in S_t} \exp[\frac{-\beta E(\omega', s')}{RT}]$. This normalization enforces that reproduction probability of a sequence-structure pair is weighted against the Boltzmann weight of the entire population. $\beta$ is the selection coefficient that takes the value $\beta = -1$ in our simulations. $R = 0.000\,198\,71\ \mathrm{kcal\,mol^{-1}\,K^{-1}}$ and $T = 310.15K$ are the Boltzmann constant and standard temperature respectively.

When a sequence is selected for replication, the child sequence is formed by copying each nucleotide of the parent RNA with an error rate of $\mu$ known as the mutation rate. $\mu$ defines the probability of incorrectly copying a nucleotide and instead randomly sampling one of the other 3 bases in $\mathbb{B}$.

#### 4.1.3 Controlling population GC content

There are two obstacles to maintaining evolving populations within the desired GC content range of $\pm 0.1$. First, an initial population of random sequences sampled uniformly from the full alphabet naturally tends converge to a GC content of 0.5. To avoid this, we sample from the alphabet with probability of sampling GC and AU equal to the desired GC content. This way our initial population has the desired nucleotide distribution. Second, when running the simulation, random mutations are able to move replicating sequences outside of the desired range, especially at extremes of mutation rate and GC content. To avoid this drift, at the selection stage, we do not select mutations that would take the sequence outside of this range. Instead, if a mutation takes a replicating sequence outside the GC range, we simply repeat the mutation process on the sequence until the child sequence has the appropriate GC content (See **Alg. 1**). Given that populations are initialized in the appropriate GC range, we are likely to find valid mutants

relatively quickly and always avoid drifting away from the target GC.

**input :** *parentSeq*, *targetGC*

**output:** *childSeq*

*childSeq* ← `mutate(`*parentSequence*`)`

**while** `computeGC(`*childSequence*`)` *not in targetGC* $\pm 0.1$ **do**
| *childSeq* ← `mutate(`*parentSeq*`)`

**end**

**return** *childSeq*

**Algorithm 1:** GC content maintaining replication

### 4.2 `RNAmutants`

The evolutionary algorithm is similar to a local search. At every time step new sequences are close to the previous population and in particular to the elements with higher *fitness*.

In contrast `RNAmutants`[36] can sample pairs of sequence-structure $(\omega, s)$ such that (1) the sequence is a $k$-mutant from a given seed $\omega_0$—for any $k$—and (2) the probability of seeing the pair is proportional to its *fitness* compared to all pairs $(\omega', s')$ where $w'$ is also an $k$-mutant of $\omega_0$.

In addition `RNAmutants` provides an unbiased control of the samples GC content allowing direct comparisons with `mateRNAl`.

We note that although the structure sampled is not in general the MFE replacing them by it does not significantly change the results, as shown in Fig. S1. Therefore we replace the sampled structure with the MFE to simplify the study.

Fo each GC content in $\{0.1, 0.3, 0.5, 0.7, 0.9\}(\pm 0.1)$ we generated 20 random seed of length 50. For each seed, at each mutational distance (i.e. number of mutations from the seed) from 0 to 50, at least 10 000 sequence-structure pairs within the target GC content of the seed were sampled from the Boltzmann distribution. The software was run on Dual Intel Westmere EP Xeon X5650 (6-core, 2.66 GHz, 12MB Cache, 95W) on the Guillimin High Performance Computing Cluster of Calcul Québec. It took over 12 000 CPU hours to complete the sampling.

### 4.2.1 Sequence-structure pairs weighted sampling

Given a seed sequence $\omega_0$, a fixed number of mutations $k$, and the ensemble $\mathbb{S}^k_{\omega_0}$ of all pairs sequence-structure such that the sequences are at hamming $k$ from $\omega_0$. Similarly to Sec. 4.1.2 the *fitness* of a sequence-structure pair $(\omega, s) \in \mathbb{S}^k_{\omega_0}$ will be its Boltzmann weight, a function of its energy. Formally, if the energy of the sequence $\omega$ in conformation $s$ is $E(\omega, s)$ then the weight of the pair is:

$$e^{\frac{-E(\omega,s)}{RT}}$$

where as before the Boltzmann constant $R$ equals $0.000\,198\,71$ $\mathrm{kcal\,mol^{-1}\,K^{-1}}$and the temperature $T$ is set at $310.15K$. The normalization factor, or partition function, $\mathcal{Z}$ can now be defined as:

$$\mathcal{Z} = \sum_{(\omega',s')\in\mathbb{S}^k_{\omega_0}} e^{\frac{-E(\omega',s')}{RT}}$$

and thus the probability of sampling a pair $(\omega, s)$ is:

$$\mathbb{P}(\omega, s) = \frac{e^{\frac{-E(\omega,s)}{RT}}}{\mathcal{Z}}.$$

By increasing $k$ from 1 to $|\omega_0|$ an exploration of whole mutational landscape of $\omega_0$ is performed. To compute $\mathcal{Z}$ for each value of $k$, `RNAmutants` has a complexity of $\mathcal{O}(n^3k^2)$. This has to be done only once per seed. The weighted sampling of the sequences themselves has complexity of $\mathcal{O}(n^2)$.

### 4.2.2 Controlling samples GC content

Due to the deep correlation between the GC content of the sequence and its energy, the GC base pair being the most energetic in the Turner model[51] which is used by `RNAmutants`, sampling from any ensemble $\mathbb{S}$ will be highly biased towards sequences with high GC content. To get a sample $(\omega, s)$ at a specific target GC content, a natural approach is to continuously sample and reject any sequence not fitting the requirements. Such an approach can yield an exponential time so a technique developed in[37] is applied.

An unbiased sampling of pairs $(\omega, s)$ for any given GC target can be obtained by modifying the

Boltzmann weights of any element $(\omega, s)$ with a term $\mathbf{w}^{\omega} \in [0,1]$ which depends on the GC content of $\omega$. At its simplest, it can be the proportion of GC in $\omega$. The weight of $(\omega, s)$ becomes

$$\mathbf{w}^{\omega} e^{\frac{-E(\omega,s)}{RT}}$$

which implies that a new partition function $\mathcal{Z}^{\mathbf{w}}$ needs to be defined as follows:

$$\mathcal{Z}^{\mathbf{w}} = \sum_{(\omega',s') \in \mathbb{S}^k_{\omega_0}} \mathbf{w}^{\omega} e^{\frac{-E(\omega',s')}{RT}}.$$

To find the weights $\mathbf{w}$ for any target GC an exact solution could be found but in practice an efficient solution consists in applying a bisection algorithm to $\mathbf{w}$. The general idea is to sample a limited number of sequences. Keep those with the desired GC content and then update $\mathbf{w}$ upwards if the average GC of the samples is lower than the target, and downwards else. In practice, only a couple of iterations are required.

### 4.3 Sequence divergence in an random replication model

We estimate the expected number of mutations in randomly replicated sequences (section 2.4) using the transition matrix defined by K. Tamura[39]. We use a mutation rate $\alpha = 0.02$ mirroring the mutation rate used in `mateRNAl`, and assume that transition and transversion rates are identical. The target GC content is represented with the variable $\theta = \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The transition matrix is shown below.

| | A | U | C | G |
|---|---|---|---|---|
| A | $1-\alpha(1+\theta)$ | $(1-\theta)\alpha$ | $\theta\alpha$ | $\theta\alpha$ |
| U | $(1-\theta)\alpha$ | $1-\alpha(1+\theta)$ | $\theta\alpha$ | $\theta\alpha$ |
| C | $(1-\theta)\alpha$ | $(1-\theta)\alpha$ | $1-(1-\alpha(1+\theta))$ | $\theta\alpha$ |
| G | $(1-\theta)\alpha$ | $(1-\theta)\alpha$ | $\theta\alpha$ | $1-(1-\alpha(1+\theta))$ |

This matrix gives us the transition rate from one generation to the next one. To obtain the mutation probabilities at the $k^{th}$ generation, we calculate the $k^{th}$ exponent of this matrix. Then, we sum the values along the main diagonal to estimate the probability of a nucleotide to be the same at the initial and $k^{th}$ generation.

## 5 Authors contributions

CO, VR, and JW designed the research, analyzed the results, and wrote the manuscript. CO and VR conducted the computational experiments.

## References


[1] Walter Gilbert. Origin of life: The RNA world. *Nature*, 319:618, February 1986. doi: 10.1038/319618a0.

[2] S R Eddy. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet*, 2(12):919–29, Dec 2001. doi: 10.1038/35103511.

[3] David P Horning and Gerald F Joyce. Amplification of RNA by an RNA polymerase ribozyme. *Proc Natl Acad Sci U S A*, 113(35):9786–91, Aug 2016. doi: 10.1073/pnas.1610103113.

[4] Sidney Becker, Ines Thoma, Amrei Deutsch, Tim Gehrke, Peter Mayer, Hendrik Zipse, and Thomas Carell. A high-yielding, strictly regioselective prebiotic purine nucleoside formation pathway. *Science*, 352(6287):833–6, May 2016. doi: 10.1126/science.aad2808.

[5] Paul G Higgs and Niles Lehman. The RNA World: molecular cooperation at the origins of life. *Nat Rev Genet*, 16(1):7–17, Jan 2015. doi: 10.1038/nrg3841.

[6] C Reidys, P F Stadler, and P Schuster. Generic properties of combinatory maps: neutral networks of RNA secondary structures. *Bull Math Biol*, 59(2):339–97, Mar 1997.

[7] Eric P Nawrocki, Sarah W Burge, Alex Bateman, Jennifer Daub, Ruth Y Eberhardt, Sean R Eddy, Evan W Floden, Paul P Gardner, Thomas A Jones, John Tate, and Robert D Finn. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res*, 43(Database issue): D130–7, Jan 2015. doi: 10.1093/nar/gku1063.

[8] Ronny Lorenz, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA package 2.0. *Algorithms Mol Biol*, 6:26, 2011. doi: 10.1186/1748-7188-6-26.

[9] Walter Fontana, Thomas Griesmacher, Wolfgang Schnabl, Peter F Stadler, and Peter Schuster. Statistics of landscapes based on free energies, replication and degradation rate constants of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 122(10):795–819, 1991.

[10] Peter Schuster and Peter F Stadler. Landscapes: Complex optimization problems and biopolymer structures. *Computers & chemistry*, 18(3):295–324, 1994.

[11] Walter Gruner, Robert Giegerich, Dirk Strothmann, Christian Reidys, Jacqueline Weber, Ivo L Hofacker, Peter F Stadler, Peter Schuster, et al. Analysis of RNA sequence structure maps by exhaustive enumeration. *Monatsh Chem*, 127(355), 1996.

[12] Peter Schuster and Walter Fontana. Chance and necessity in evolution: lessons from RNA. *Physica D: Nonlinear Phenomena*, 133(1-4):427–452, September 1999. ISSN 01672789. doi: 10.1016/S0167-2789(99)00076-7. URL `http://linkinghub.elsevier.com/retrieve/pii/S0167278999000767`.

[13] Peter Schuster. Evolution in silico and in vitro: the RNA model. *Biological chemistry*, 382(9): 1301–1314, 2001.

[14] E van Nimwegen, J P Crutchfield, and M Huynen. Neutral evolution of mutational robustness. *Proc Natl Acad Sci U S A*, 96(17):9716–20, Aug 1999.

[15] L W Ancel and W Fontana. Plasticity, evolvability, and modularity in RNA. *J Exp Zool*, 288 (3):242–83, Oct 2000.

[16] C O Wilke. Selection for fitness versus selection for robustness in RNA secondary structure folding. *Evolution; international journal of organic evolution*, 55(12):2412–20, December 2001. ISSN 0014-3820. URL `http://www.ncbi.nlm.nih.gov/pubmed/11831657`.

[17] Jacobo Aguirre, Javier M Buldú, Michael Stich, and Susanna C Manrubia. Topological structure of the space of phenotypes: the case of RNA neutral networks. *PLoS One*, 6(10):e26324, 2011. doi: 10.1371/journal.pone.0026324.

[18] Kamaludin Dingle, Steffen Schaper, and Ard A Louis. The structure of the genotype-phenotype map strongly constrains the evolution of non-coding RNA. *Interface Focus*, 5(6):20150053, Dec 2015. doi: 10.1098/rsfs.2015.0053.

[19] Anne Kupczok and Peter Dittrich. Determinants of simulated RNA evolution. *Journal of theoretical biology*, 238(3):726–35, February 2006. ISSN 0022-5193. doi: 10.1016/j.jtbi.2005.06.019. URL `http://www.ncbi.nlm.nih.gov/pubmed/16098538`.

[20] Michael Stich, Carlos Briones, and Susanna C Manrubia. Collective properties of evolving molecular quasispecies. *BMC evolutionary biology*, 7:110, January 2007. ISSN 1471-2148. doi: 10.1186/1471-2148-7-110. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1934359\&tool=pmcentrez\&rendertype=abstract`.

[21] Michael Stich, Susanna C Manrubia, and Ester La. Variable Mutation Rates as an Adaptive Strategy in Replicator Populations. *PLoS ONE*, 5(6), 2010. doi: 10.1371/Citation.

[22] Michael Stich, Carlos Briones, and Susanna C Manrubia. On the structural repertoire of pools of short, random RNA sequences. *Journal of theoretical biology*, 252(4):750–763, 2008.

[23] P Schuster, W Fontana, P F Stadler, and I L Hofacker. From sequences to shapes and back: a case study in RNA secondary structures. *Proc Biol Sci*, 255(1344):279–84, Mar 1994. doi: 10.1098/rspb.1994.0040.

[24] W Fontana and P Schuster. Shaping space: the possible and the attainable in RNA genotype-phenotype mapping. *J Theor Biol*, 194(4):491–515, Oct 1998. doi: 10.1006/jtbi.1998.0771.

[25] M Levy and A D Ellington. The descent of polymerization. *Nat Struct Biol*, 8(7):580–2, Jul 2001. doi: 10.1038/89601.

[26] Eric J Hayden and Niles Lehman. Self-assembly of a group I intron from inactive oligonucleotide fragments. *Chem Biol*, 13(8):909–18, Aug 2006. doi: 10.1016/j.chembiol.2006.06.014.

[27] Nilesh Vaidya, Michael L Manapat, Irene A Chen, Ramon Xulvi-Brunet, Eric J Hayden, and Niles Lehman. Spontaneous network formation among cooperative RNA replicators. *Nature*, 491(7422):72–7, Nov 2012. doi: 10.1038/nature11549.

[28] A A Beaudry and G F Joyce. Directed evolution of an RNA enzyme. *Science*, 257(5070): 635–41, Jul 1992.

[29] D P Bartel and J W Szostak. Isolation of new ribozymes from a large pool of random sequences. *Science*, 261(5127):1411–8, Sep 1993.

[30] Erik A Schultes, Alexander Spasic, Udayan Mohanty, and David P Bartel. Compact and ordered collapse of randomly generated RNA sequences. *Nat Struct Mol Biol*, 12(12):1130–6, Dec 2005. doi: 10.1038/nsmb1014.

[31] K Salehi-Ashtiani and J W Szostak. In vitro evolution suggests multiple origins for the hammerhead ribozyme. *Nature*, 414(6859):82–4, Nov 2001. doi: 10.1038/35102081.

[32] Péter Szabó, István Scheuring, Tamás Czárán, and Eörs Szathmáry. In silico simulations reveal that replicators with limited dispersal evolve towards higher efficiency and fidelity. *Nature*, 420(6913):340–3, Nov 2002. doi: 10.1038/nature01187.

[33] Carlos Briones, Michael Stich, and Susanna C Manrubia. The dawn of the RNA World: toward functional complexity through ligation of random RNA oligomers. *RNA*, 15(5):743–9, May 2009. doi: 10.1261/rna.1488609.

[34] Julie A Shay, Christopher Huynh, and Paul G Higgs. The origin and spread of a cooperative replicase in a prebiotic chemical system. *J Theor Biol*, 364:249–59, Jan 2015. doi: 10.1016/j.jtbi.2014.09.019.

[35] Nikola A Ivica, Benedikt Obermayer, Gregory W Campbell, Sudha Rajamani, Ulrich Gerland, and Irene A Chen. The paradox of dual roles in the RNA world: resolving the conflict between stable folding and templating ability. *Journal of molecular evolution*, 77(3):55–63, 2013.

[36] Jérôme Waldispühl, Srinivas Devadas, Bonnie Berger, and Peter Clote. Efficient algorithms for probing the RNA mutation landscape. *PLoS Comput Biol*, 4(8):e1000124, Aug 2008. doi: 10.1371/journal.pcbi.1000124.

[37] Jérôme Waldispühl and Yann Ponty. An unbiased adaptive sampling algorithm for the exploration of RNA mutational landscapes under evolutionary pressure. *Journal of Computational Biology*, 18(11):1465–1479, 2011.

[38] Matthew C Cowperthwaite, Evan P Economo, William R Harcombe, Eric L Miller, and Lauren Ancel Meyers. The ascent of the abundant: how mutational networks constrain evolution. *PLoS Comput Biol*, 4(7):e1000110, Jul 2008. doi: 10.1371/journal.pcbi.1000110.

[39] K Tamura. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol*, 9(4):678–87, Jul 1992.

[40] Paul P Gardner, Barbara R Holland, Vincent Moulton, Mike Hendy, and David Penny. Optimal alphabets for an RNA world. *Proc Biol Sci*, 270(1520):1177–82, Jun 2003. doi: 10.1098/rspb.2003.2355.

[41] Sandra Smit, Michael Yarus, and Rob Knight. Natural selection is not required to explain universal compositional patterns in rRNA secondary structure categories. *RNA*, 12(1):1–14, Jan 2006. doi: 10.1261/rna.2183806.

[42] Alex Levin, Mieszko Lis, Yann Ponty, Charles W O'Donnell, Srinivas Devadas, Bonnie Berger, and Jérôme Waldispühl. A global sampling approach to designing and reengineering RNA secondary structures. *Nucleic Acids Res*, 40(20):10041–52, Nov 2012. doi: 10.1093/nar/gks768.

[43] S F Greenbury and S E Ahnert. The organization of biological sequences into constrained and unconstrained parts determines fundamental properties of genotype-phenotype maps. *J R Soc Interface*, 12(113):20150724, Dec 2015. doi: 10.1098/rsif.2015.0724.

[44] Susanna Manrubia and José A Cuesta. Distribution of genotype network sizes in sequence-to-structure genotype-phenotype maps. *J R Soc Interface*, 14(129), Apr 2017. doi: 10.1098/rsif.2016.0976.

[45] Robert Pascal, Addy Pross, and John D Sutherland. Towards an evolutionary theory of the origin of life based on kinetics and thermodynamics. *Open Biol*, 3(11):130156, Nov 2013. doi: 10.1098/rsob.130156.

[46] A R Hill, Jr, L E Orgel, and T Wu. The limits of template-directed synthesis with nucleoside-5'-phosphoro(2-methyl)imidazolides. *Orig Life Evol Biosph*, 23(5-6):285–90, Dec 1993.

[47] J Waldispühl, B Behzadi, and J-M Steyaert. An approximate matching algorithm for finding (sub-)optimal sequences in S-attributed grammars. *Bioinformatics*, 18 Suppl 2:S250–9, 2002.

[48] Susanna C Manrubia and Carlos Briones. Modular evolution and increase of functional complexity in replicating RNA molecules. *RNA*, 13(1):97–107, 2007.

[49] Sudha Rajamani, Justin K Ichida, Tibor Antal, Douglas A Treco, Kevin Leu, Martin A Nowak, Jack W Szostak, and Irene A Chen. Effect of stalling after mismatches on the error catastrophe in nonenzymatic nucleic acid replication. *J Am Chem Soc*, 132(16):5880–5, Apr 2010. doi: 10.1021/ja100780p.

[50] Eörs Szathmáry. The origin of replicators and reproducers. *Philos Trans R Soc Lond B Biol Sci*, 361(1474):1761–76, Oct 2006. doi: 10.1098/rstb.2006.1912.

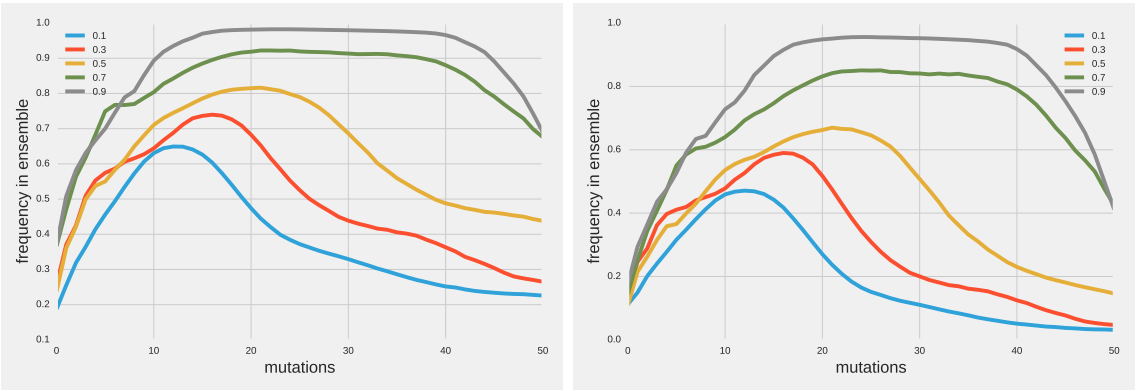
[51] Douglas H Turner and David H Mathews. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic acids research*, page gkp892, 2009.

**Supplemental Materials: "The necessary emergence of structural complexity in self-replicating RNA populations"**

Carlos Oliver, Vladimir Reinharz, Jérôme Waldispühl

# 1 Sampling schemes

`RNAmutants` samples sequence-structure pairs according to their ensemble weight. This method therefore does not guarantee the sampling of the minimum free energy (MFE) structure for a given sequence. However, for a given "suboptimal" sampling we can use `RNAfold` to obtain the MFE structure for every sequence, producing a set of sequence-structure pairs that can be directly compared to `mateRNAl` data which samples only MFE structures. In **Fig. S1** we show that this procedure produces sequence-structure pairs with nearly identical ensemble frequencies, allowing us to proceed with MFE structures in downstream analysis.




(a) MFE (b) Sampled

Figure S1: Frequency of sequence-structure pair in Boltzmann ensemble for MFE and suboptimal sampling in RNAmutants.

# 2 Energy landscapes

With free energy being the driving force in our simulations, we next compared mean energy values for mutation populations in `mateRNAl` and `RNAmutants`. We show in **Fig. S2** that a random starting seed in `RNAmutants` is on average under 15 mutations away from the global minimum of its mutational ensemble. This feature of the energy of mutational landscapes is exploited by the

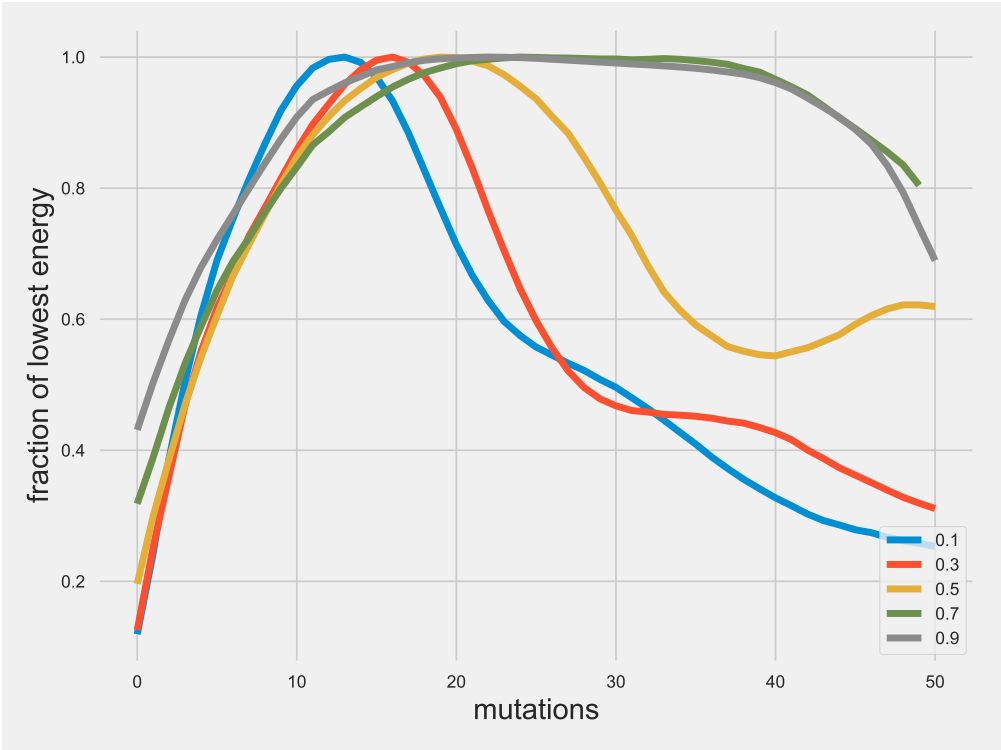
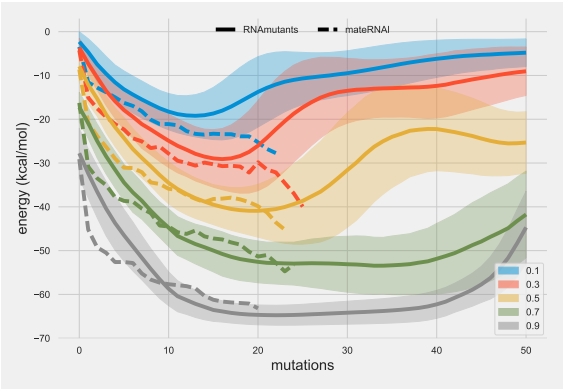



Figure S2: Percentage of the global minimum energy reached by RNAmutants by mutational distance.

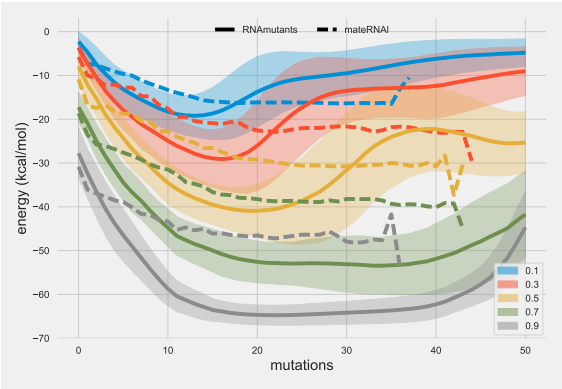
local search nature of `mateRNAl` which is able to rapidly identify stable solutions with very few mutations. More specifically, we tested various mutation rates ($\mu$) for `mateRNAl` and superimposed mean values to those obtained by `RNAmutants`, **Fig. S3**. We note that higher mutation rates lead to deeper mutational explorations yet never reaching as far as `RNAmutants`. Regardless, it appears that the most stable structures are maintained at lower mutation rates, suggesting a tradeoff between exploration and refinement. **Fig. S4a** shows an example of population energy over generation time instead of over mutation bins again denoting the rapid adaptation behaviour. Finally, we note that although mutation rate has an important effect on the depth of evolutionary searches, it appears that it has a only a slight effect on the global mean energy for a given simulation **Fig. S4b** which is in agreement with the findings of mutational networks by `RNAmutants`.

## 3 Structural complexity

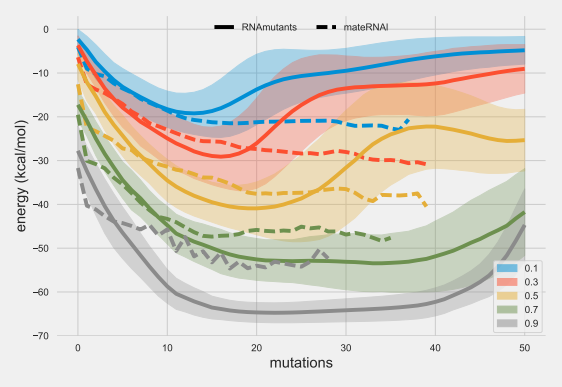
Here we summarize the structural features discovered by `mateRNAl` and `RNAmutants`. In this study we distinguish between three major structural features: stack, internal loop, and multiloop. The




(a) $\mu = 0.001$

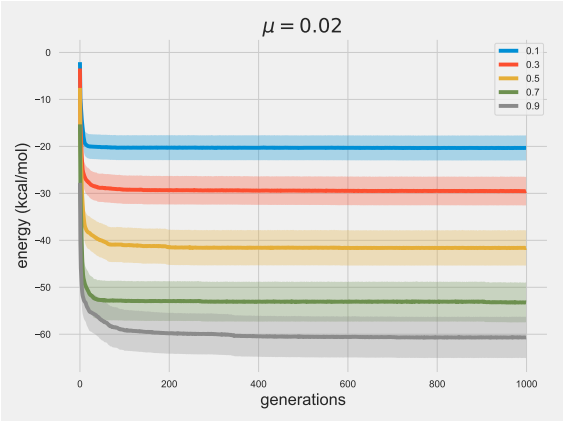



(b) $\mu = 0.1$

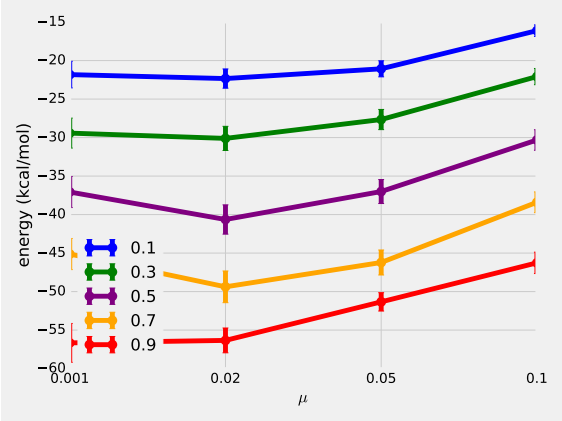



(c) $\mu = 0.05$

Figure S3: Average energy comparison between `mateRNAl` and `RNAmutants` using different mutation rates for `mateRNAl`.




(a) Population energy vs. number of generations.




(b) Effect of mutation rate and GC content on energy optimization.

Figure S4: Average energy for `mateRNAl` simulations. **S4a** is a sample run showing energy over generation time. **S4b** shows the mean energy over the entire simulation for all mutation rates and GC contents.

stack is produced from consecutive base pairing interactions and thus higher stack numbers form longer stem structures. These represent the simplest of the structural motifs. The internal loop occurs when a stacking is interrupted by unpaired bases forming a loop like structure within a stem. A multi-loop is also an unpaired region that forms a loop but which connects three or more stems, also known as a junction. **Fig. S5** shows the effect of GC content on the occurence of internal looping structures as a function of mutational distance in `RNAmutants` simulations. In **Fig. S6** we show the effect of GC content and mutation rate on all three structural motifs and show that mutation rate appears to have a strong impact on the discovery of diverse motifs. However, for motifs such as multiloops, the absolute frequencies observed are too low to make any statistically sound claims. Global means for all structural motifs in `RNAmutants` and `mateRNAl` are summarized in **Table S1** where we can see the enrichment of complex motifs in `RNAmutants` simulations compared to `mateRNAl`.

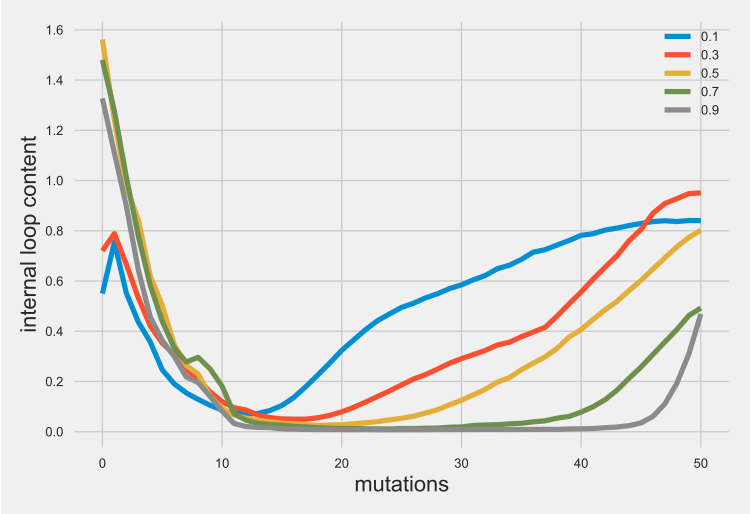
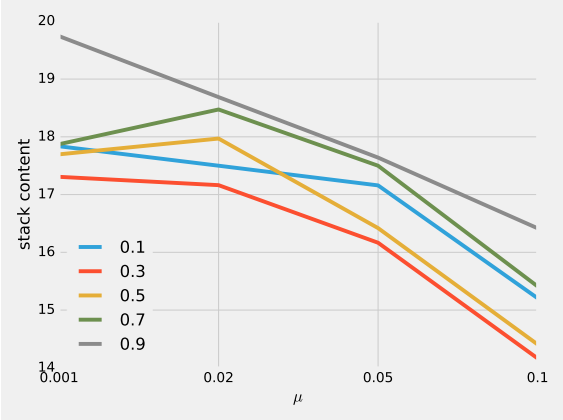


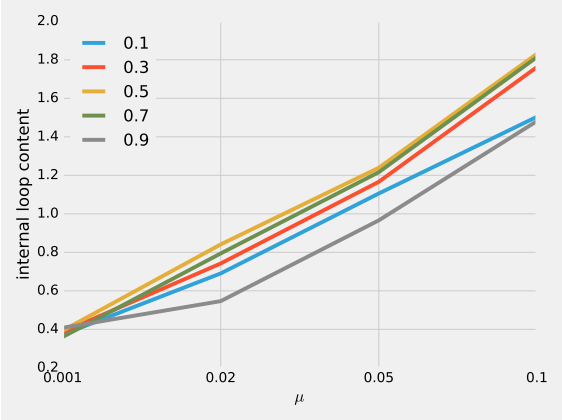

Figure S5: Fraction of $k$-mutant neighbourhoods populated with internal looping structures in `RNAmutants`.

## 4 Analysis of Hamming neighbourhoods

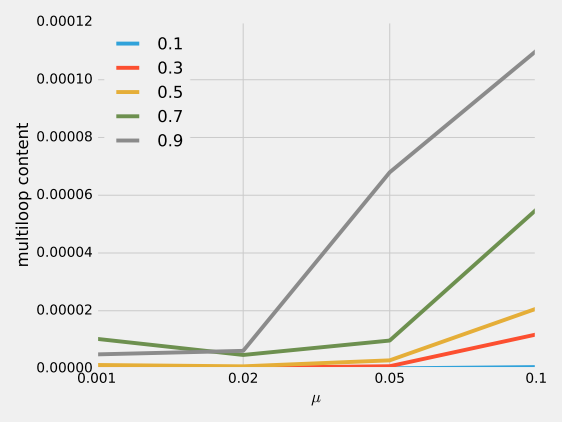
We provide complementary data to characterize the sequences and structures available at specific mutational distances and GC contents. **Fig. S7** indicates the number of sequences with exactly $k$ mutations and with a GC content within $\pm 0.1$ of the target GC content. Thus, it shows the size of the neighbourhoods.




(a) Stacking pair content




(b) Internal loop content




(c) Multiloop content

Figure S6: Effect of mutation rate and GC content on structural complexity of structures found with `mateRNA1`.

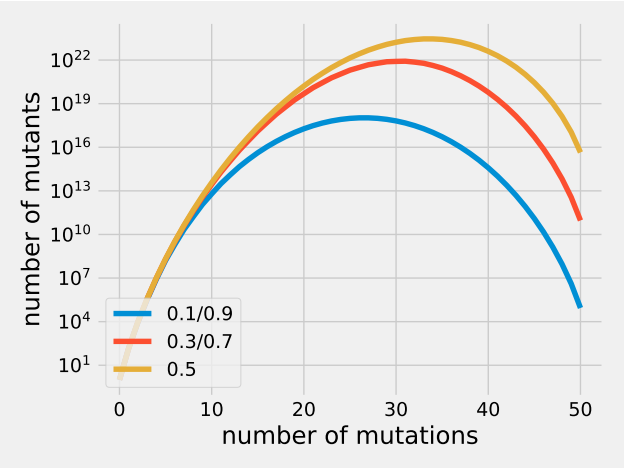



Figure S7: Number of sequences in hamming neighbourhoods.

| GC | stack avg | stack std | internal avg | internal std | multi avg | multi std |
|---|---|---|---|---|---|---|
| 0.1 | 13.977 | 3.18763 | 0.493055 | 0.264608 | 2.91143e-06 | 6.48634e-06 |
| 0.3 | 13.3715 | 3.41882 | 0.383487 | 0.286511 | 0.000710177 | 0.000804852 |
| 0.5 | 15.866 | 3.23955 | 0.321108 | 0.331889 | 0.00330513 | 0.00374271 |
| 0.7 | 19.4817 | 2.63143 | 0.204601 | 0.320884 | 0.00057179 | 0.000703733 |
| 0.9 | 20.5086 | 2.1998 | 0.142698 | 0.286028 | 5.17749e-05 | 0.000150411 |

(a) `RNAmutants` secondary structure content

| GC | stack avg | stack std | internal avg | internal std | multi avg | multi std |
|---|---|---|---|---|---|---|
| 0.1 | 17.8367 | 0.367748 | 0.372504 | 0.0209848 | 0 | 0 |
| 0.3 | 17.3073 | 0.606713 | 0.385716 | 0.0262441 | 0 | 0 |
| 0.5 | 17.6965 | 0.459124 | 0.393226 | 0.0637715 | 1.21878e-06 | 3.85605e-05 |
| 0.7 | 17.872 | 0.432849 | 0.36015 | 0.0604876 | 1.0273e-05 | 0.00011642 |
| 0.9 | 19.7432 | 0.596439 | 0.408057 | 0.0598234 | 4.8869e-06 | 0.000125136 |

(b) `mateRNAl` $\mu = 0.001$

| GC | stack avg | stack std | internal avg | internal std | multi avg | multi std |
|---|---|---|---|---|---|---|
| 0.1 | 15.2058 | 0.277693 | 1.50402 | 0.0384733 | 4.89541e-07 | 3.12498e-06 |
| 0.3 | 14.1637 | 0.293036 | 1.76198 | 0.0462676 | 1.17807e-05 | 2.36637e-05 |
| 0.5 | 14.4054 | 0.267386 | 1.83133 | 0.028948 | 2.07133e-05 | 3.29296e-05 |
| 0.7 | 15.4102 | 0.243479 | 1.81397 | 0.0446219 | 5.49919e-05 | 0.000107943 |
| 0.9 | 16.4166 | 0.255144 | 1.48154 | 0.121822 | 0.000109995 | 0.000136078 |

(c) `mateRNAl` $\mu = 0.1$

Table S1: Summary of global means for stack, internal loop and multiloop content in `mateRNAl` and `RNAmutants` for low ($\mu = 0.001$) and high ($\mu = 0.1$)mutation rates.