On the emergence of structural complexity in RNA replicators

CARLOS G. OLIVER,[1] VLADIMIR REINHARZ,[2] and JÉRÔME WALDISPÜHL[1]

[1]School of Computer Science, McGill University, Montreal, QC H3A 2B3, Canada
[2]Center for Soft and Living Matter, Institute for Basic Science, Ulsan 34126, South Korea

ABSTRACT

The RNA world hypothesis relies on the ability of ribonucleic acids to spontaneously acquire complex structures capable of supporting essential biological functions. Multiple sophisticated evolutionary models have been proposed for their emergence, 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 multibranched 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-nt-long RNA sequences. Our results reveal that the sequence landscape with stable structures is enriched with multibranched structures at a length scale coinciding with the appearance of complex structures in RNA databases. A random replication mechanism preserving a 50% GC content may suffice to explain a natural enrichment of stable complex structures in populations of functional RNAs. In contrast, an evolutionary mechanism eliciting the most stable folds at each generation appears to help reaching multibranched structures at highest GC content.

Keywords: RNA world; evolution; multiloops; self-replication

INTRODUCTION

RNA are versatile molecules that can fulfill virtually all fundamental needs and functions of the living, from storing information to catalyzing chemical reactions and regulating gene expression. The RNA world hypothesis (Gilbert 1986; Eddy 2001) builds upon this observation to describe a scenario of the emergence of life based on RNAs. Despite criticisms (Orgel 2004; Shapiro 2007; Robertson and Joyce 2012), recent studies presented plausible paths toward an early assembly of RNA molecules, which has contributed to strengthening this hypothesis (Powner et al. 2009; Ritson and Sutherland 2012; Becker et al. 2016; Horning and Joyce 2016; Pearce et al. 2017). Yet, the emergence of nucleic acids is only one part of this problem. We also need to elucidate the mechanisms that enabled the discovery of functional molecules and the transmission of genetic information (Szostak 2012).

Many noncoding RNAs acquire functions through structures. Classically, we describe these structures at two levels of abstraction. The secondary structure encompasses the Watson–Crick and wobble base pairs, while the tertiary structure describes the 3D coordinates of all atoms. Noticeably, RNA secondary structures are more conserved than sequences and thus provide a reliable signature of RNA function (Nawrocki et al. 2015). Moreover, many essential molecular functions are supported by nucleic acids with complex shapes often characterized by the occurrence of a *k*-way junction (a loop connecting three or more stem-like regions also known as multiloop (ML); see Fig. 1C) in their secondary structure (e.g., 5s rRNA, tRNA, hammerhead ribozyme). A theory describing how RNA populations evolve to "discover" these functional multibranched secondary structures is, therefore, an important step toward a validation of the RNA world hypothesis (Higgs and Lehman 2015).

Since the first analysis of RNA neutral networks (i.e., networks of RNA sequences with identical structures) (Reidys et al. 1997), computer simulations are the method of choice for characterizing the evolutionary landscape and population dynamics of structured RNA molecules. Indeed, secondary structures can be reliably predicted from sequence data only (Lorenz et al. 2011), allowing

Corresponding author: jeromew@cs.mcgill.ca
Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.070391.119.

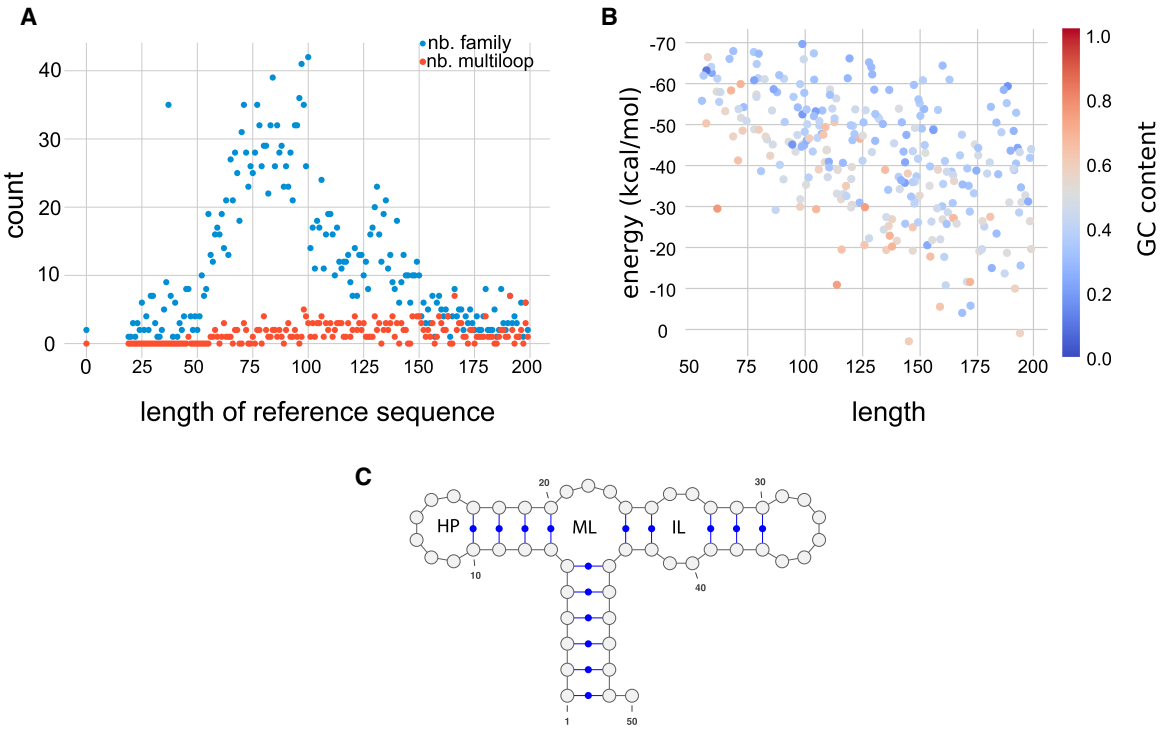

FIGURE 1. Statistics on Rfam families (Nawrocki et al. 2015). (*A*) We plot the number of families with respect to the average length of the sequences in these families. Red dots show the numbers of families with a consensus structure that contains a ML, while blue dots show those without. (*B*) We plot the average folding energy and length of sequences for each Rfam family having multibranched consensus structures. The color indicates the average GC content of the family. (*C*) An illustration of an RNA 2D fold (generated using VARNA [Darty et al. 2009]) with its secondary structure elements labeled as internal loop (IL), ML, and hairpin (HP).

fast and accurate prediction of phenotypes (i.e., secondary structure) from genotypes (i.e., sequence).

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 Schuster and coworkers set up the basis of a theoretical framework to study the 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) (Fontana et al. 1991; Schuster and Stadler 1994; Gruner et al. 1996; Reidys et al. 1997; Schuster and Fontana 1999; Schuster 2001). This work inspired numerous computational studies that refined our understanding of neutral models (van Nimwegen et al. 1999; Ancel and Fontana 2000; Wilke 2001; Aguirre et al. 2011; Dingle et al. 2015), as well as kinetics of populations of evolving nucleic acids (Kupczok and Dittrich 2006; Stich et al. 2007, 2010).

In this paper, we perform computer simulations to study the evolutionary mechanisms that enabled the emergence of multibranched secondary structures. This feature turns out to be relatively common even for short RNAs. An analysis of the consensus secondary structures available in the Rfam database reveals that MLs can be found in ~10% of RNA families whose average size of sequences ranges from 50 to 100 nt (see Fig. 1A). This observation contrasts with earlier studies that revealed that the vast majority of predicted minimum free energy (MFE) secondary structures obtained from a uniform sampling of shorter RNA sequences (i.e., 35 nt) are stem-like structures (i.e., no ML) (Stich et al. 2008), which render a spontaneous emergence of complex structures (i.e., secondary structures with a ML) unlikely on such short sequences.

Nonetheless, computational studies of RNA sequence–structure maps showed that neutral networks percolate the whole sequence landscape (Schuster et al. 1994; Fontana and Schuster 1998). 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 structures.

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 (Higgs and Lehman 2015). Such molecules are long (200 nt) and thus unlikely to be discovered randomly. Instead, it has been suggested that evolution proceeded in stages (Levy and Ellington 2001). Polymerases were assembled from smaller monomers (~50 nt) that are more likely to

emerge from prebiotic chemistry (Hayden and Lehman 2006; Vaidya et al. 2012). At this point, and not before, parallel natural selection processes of specific functional structures could be triggered.

Interestingly, in vitro experiments revealed the extreme versatility of random nucleic acids (Beaudry and Joyce 1992; Bartel and Szostak 1993; Schultes et al. 2005; Pressman et al. 2017), and suggested that essential RNA molecules such as the hammerhead ribozyme could have multiple origins (Salehi-Ashtiani and Szostak 2001). All together, these observations reinforce the plausibility of a spontaneous emergence of multiple functional subunits. 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 favored the birth and growth of structural complexity from replications of small monomers. Computational studies have been of tremendous help to explore various scenarios. In particular, numerical simulations enabled us to study the effects of polymerization on mineral surfaces (Szabó et al. 2002; Briones et al. 2009) or the importance of spatial diffusion (Shay et al. 2015). Noticeably, the majority of these scenarios are proposing a development of RNA structural complexity outside a cellular barrier. But this assumption results in major challenges. The first of them is to explain how a system that evolved and adapted in an exposed milieu would transition to a membrane-protected environment. In particular, the presence of a membrane would radically change the tradeoff between stability and complexity of RNA structures. Indeed, stable folds often lack the complexity necessary to support novel functions but are more resilient to harsh precellular environments (Ivica et al. 2013).

In this work, we aim to characterize the structural repertoire accessible by replicating RNA populations. This scenario is compatible with the hypothesis of an early development of membranes (Chen et al. 2005; Szostak 2012; Adamala and Szostak 2013; O'Flaherty et al. 2018) and does not require invoking hybridization on surfaces, or the presence of large self-replicating ribozymes (Paul and Joyce 2002). Importantly, we exclude directed evolution scenarios characterized by a progressive adaptation to a phenotype (e.g., replication with errors minimizing the distance of MFE structures to a target structure [Schuster 2006]). Instead, we rely on intrinsic forces (i.e., increasing secondary structure stability) for driving the evolutionary process and study the impact of GC content bias.

We apply customized algorithms to study the distribution of structures accessible to all mutant sequences with 50 nt (Waldispühl et al. 2008; Waldispühl and Ponty 2011). This approach considerably expands the scope and significance of comprehensive RNA evolutionary studies that were previously limited to sequences with <20 nt (Gruner et al. 1996; Cowperthwaite et al. 2008), or restricted to explore a small fraction of the sequence landscape of sequences (Stich et al. 2008; Dingle et al. 2015). Our simulations reveal that based on the strength of the selective pressure applied on the replicating population, different GC content biases facilitate the emergence of complex secondary structures with *k*-way junctions. In the absence of selective pressure, low to medium GC contents (0.3–0.5) may suffice to explain the distribution of ML structures observed in databases. In contrast, high GC contents help populations eliciting sequences with stable folds to discover structure with *k*-way junctions. These results provide valuable insights into previous contributions studying GC content biases in stability–flexibility tradeoffs (Leu et al. 2011), prebiotic nucleotide distributions (Penny and Poole 1999; Gardner et al. 2003) and the accessibility of complex phenotypes (Stich et al. 2008).

RESULTS

Our approach

We apply complementary techniques to explore the RNA mutation landscape and characterize the structures accessible from an initial pool of random sequences under distinct evolutionary scenarios (see Fig. 2). Importantly, our analysis explicitly models GC content bias to understand the effect of potential nucleotide scarcity in prebiotic conditions (Penny and Poole 1999; Gardner et al. 2003).

Our first algorithm `RNAmutants` (see the "`RNAmutants`" section) enumerates all mutant sequences and samples the ones with the *globally* lowest folding energy (Waldispühl et al. 2008). It enables us to identify the most stable structures accessible through mutation processes. Noticeably, `RNAmutants` sorts the sequences by the number of mutations separating them from the initial population. We note that `RNAmutants` does not constitute a model of replicating populations but rather a sampling method to obtain representative statistics on the space of stable mutant sequences.

An important element of this work is the interpretation of `RNAmutants` curves where the *x*-axis shows the number of mutations (e.g., Fig. 3). When this number of mutations is null or very low, the data represents the values we expect from uniformly sampled sequences. In contrast, large numbers of mutations (50 being the maximum number of mutations on sequences of length 50 since it is possible that all positions contain a mutation) are associated with an increased preference for mutants with stable folds and a loss of sequence specificity. The largest hamming neighborhoods (see Supplemental Fig. S7) are therefore the ones with strongest thermodynamical pressure.

Equipped with a global view of the mutational landscape from `RNAmutants`, we implemented replication-based

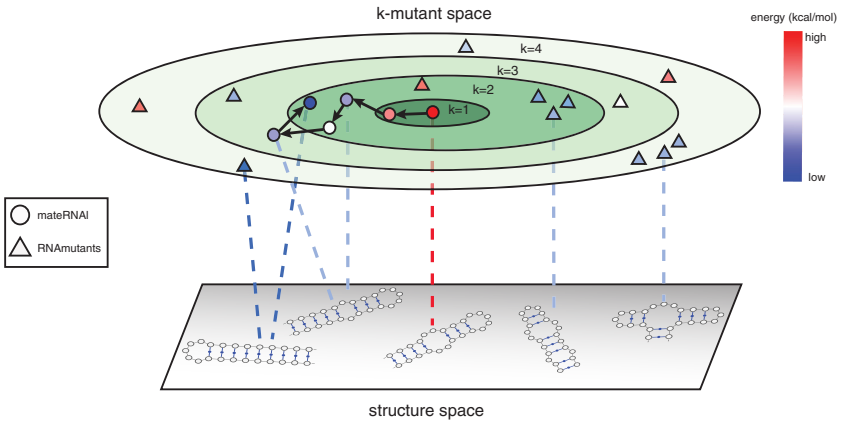

FIGURE 2. Illustration of mutational sampling methods. Concentric ellipses represent the space of sequence *k*-mutant neighborhoods around a root sequence pictured at the *center* of the ellipses. Each ring holds all sequences that are *k* mutations from the root. We show the contrast between an evolutionary 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 color of the sampled mutants denotes the energy of the sequence–structure pair sampled. In both sampling methods, sequence–structure pairs with lower energies are favored. 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`.

algorithms to understand how such a landscape could be traversed by populations. We present `mateRNAl` which has been developed for this study. `mateRNAl` simulates the evolution of a population of replicating RNA sequences that preferentially selects the most stable structures under GC content bias. We also developed variants of this algorithm enabling us to study the antagonistic effect of a negative selection pressure against the most stable structures. Since initial pools of random 50 nt sequences are unlikely to contain self-replicating RNA, we frame our experiments in the context of noncatalytic replication (Szostak 2012; O'Flaherty et al. 2019) which could have been at play until structures with self-replicating abilities are discovered (Robertson and Joyce 2014).

We study the evolutionary landscape of RNA sequences of length 50 preserving GC contents varying between 0.1 and 0.9 (1 being full G or C sequences). The size of molecules is of particular interest as 50 nt appears to be the upper limit for nonenzymatic self-replicating processes (Higgs and Lehman 2015), but also because multibranched structures occur only on RNAs with sizes above 40 (see Fig. 1). It also turns out to be the minimal known size for a natural ribozyme (Ferré-D'Amaré and Scott 2010). Shorter ribozymes have been synthesized (Turk et al. 2010) yet these would not feature complex secondary structure motifs characteristic of most ribozymes.

Energy landscape of RNA mutational networks

We start by characterizing the distribution of folding energies of stable structures accessible from random seeds in

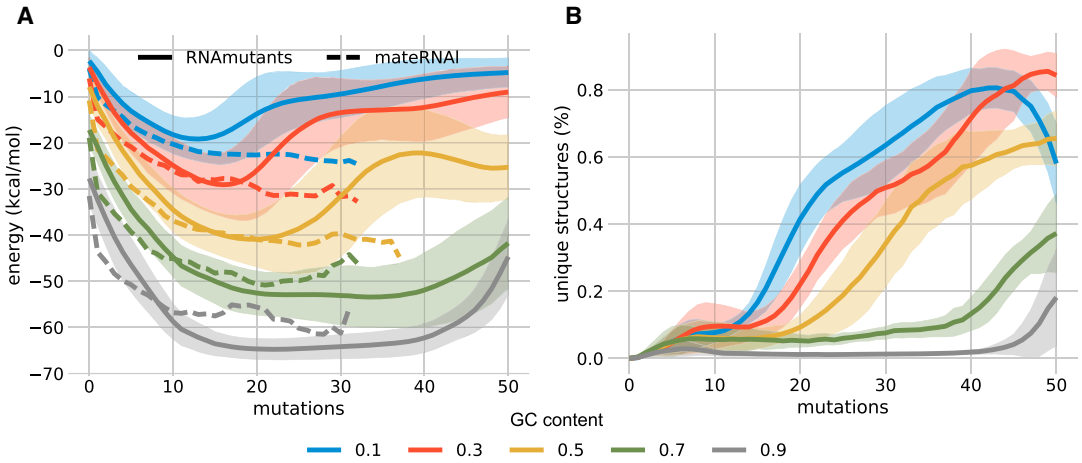

FIGURE 3. (*A*) Energy of `RNAmutants` and `mateRNAl` mutational landscape. Shaded regions include one standard deviation of `RNAmutants` energy per mutational distance. Dashed lines mark mean values for `mateRNAl` energies binned by mutations from starting sequence using mutation rate $\mu = 0.02$.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

the mutational landscape using `RNAmutants`. Our simulations show that, initially, increasing mutational distances from random seeds results in more stable structures at all GC content regimes (see solid lines in Fig. 3). We observe that the folding energies of the samples represent at least 80% of the global minimum energy attainable overall k mutations within <10 mutations from the seed (see Supplemental Fig. S3). This suggests that over short evolutionary periods, mutations can play an important stabilizing role as stable structures are likely to be found near the seed.

In 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 behavior 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, less stable sequences outnumber lower energy sequences with high Boltzmann weights (see Supplemental Fig. S2).

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

Next, we characterize the diversity of structures sampled by `RNAmutants`. First, we calculate the number of unique structures found in each mutant neighborhood. Figure 3B shows two distinct regimes. At low to intermediate GC contents ($\leq$0.5), the immediate neighborhood ($\leq$10 mutations) of seed sequences is dominated by a few stable structures, after that the percentage of unique structures suddenly increases. This sudden change of regime accompanies the increase of the average folding energies observed in Figure 3A. In contrast, at higher GC contents ($\geq$0.7) the structural diversity increases only at very large mutational distances ($\geq$40), when the footprint from the seed is almost fully lost.

We also explore the repertoire of secondary structure elements represented in the low energy mutational neighborhoods. Figure 4A and Supplemeental Figure S5 show the percentage of secondary structures with an IL or ML. Although these structural elements are relatively frequent in random sequences (i.e., seeds), they are also not very stable (Fig. 3A). When the mutational distance increases, mutations tend to create stable stem–loop structures and erase irregularities in the original phenotypes. Although, after this stabilization phase, the fraction of internal and multibranched loops rises again explaining the structural diversity discussed above.

In this work, we focus on the occurrence of multibranched secondary structures because this motif is often found in functional RNAs such as the hammerhead ribozyme. While previous studies of short randomly sampled RNA sequences (<35 nt) have shown that simple HP structures dominate the low-energy structural landscape, we find that for longer molecules (i.e., 50 nt) this landscape is enriched with complex multibranched structures. This finding is in good qualitative agreement with databases of evolved structures where ML structures emerge in families slightly under 50 nt long (Fig. 1A).

Across all runs, we sampled a total of 9419 sequences containing a ML (no more than 1 ML per structure was ever observed as is to be expected for such length scales). Interestingly, we find that unlike ILs, MLs occur under very specific conditions in our sampling. We identify a clear surge in ML frequency at a mutational distance of ~35 (see Fig. 4A), with a mean GC content of ~0.45 (see Fig. 4C). Furthermore, their energy distribution is tightly centered around −15 kcal mol^{-1} (see Fig. 4B). These values are remarkably close to those of multibranched structures of similar lengths observed in the Rfam database (see Fig. 1A). In particular, the latter shows a clear bias toward medium GC contents as we identified 148 Rfam families with MLs with a GC content of 0.5 (among all Rfam families with sequences having at most 200 nt), 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 are a general property of the distribution of MLs in the mutational landscape given the sequence entropy of the set of sequences containing MLs is quite high (0.945 out of 1). This indicates that the observed properties are likely a feature of the GC content bias and not due to overrepresentation by isolated groups of similar sequences. Further analysis carried out in the "Random replication without selection" section suggests that this enrichment of complex structures is not simply an artifact of larger Hamming neighborhoods that accompanies deeper mutational explorations.

We also note a smaller peak of ML occurrences closer to the seed sequences (~6 mutations) for higher GC contents around 0.7. Interestingly, with folding energies ranging from −25 to −40 kcal mol^{-1}, these multibranched structures are significantly more stable than those present in the main peak (see Fig. 4B). This is also in agreement

Observation 1

- Low and intermediate GC contents ($\leq$0.5) promote structural diversity.
- Internal and multibranched loops are relatively frequent in the low-energy landscape.
- The frequency of stable multibranched loops in the low-energy landscape is in good qualitative agreement with the Rfam distribution.

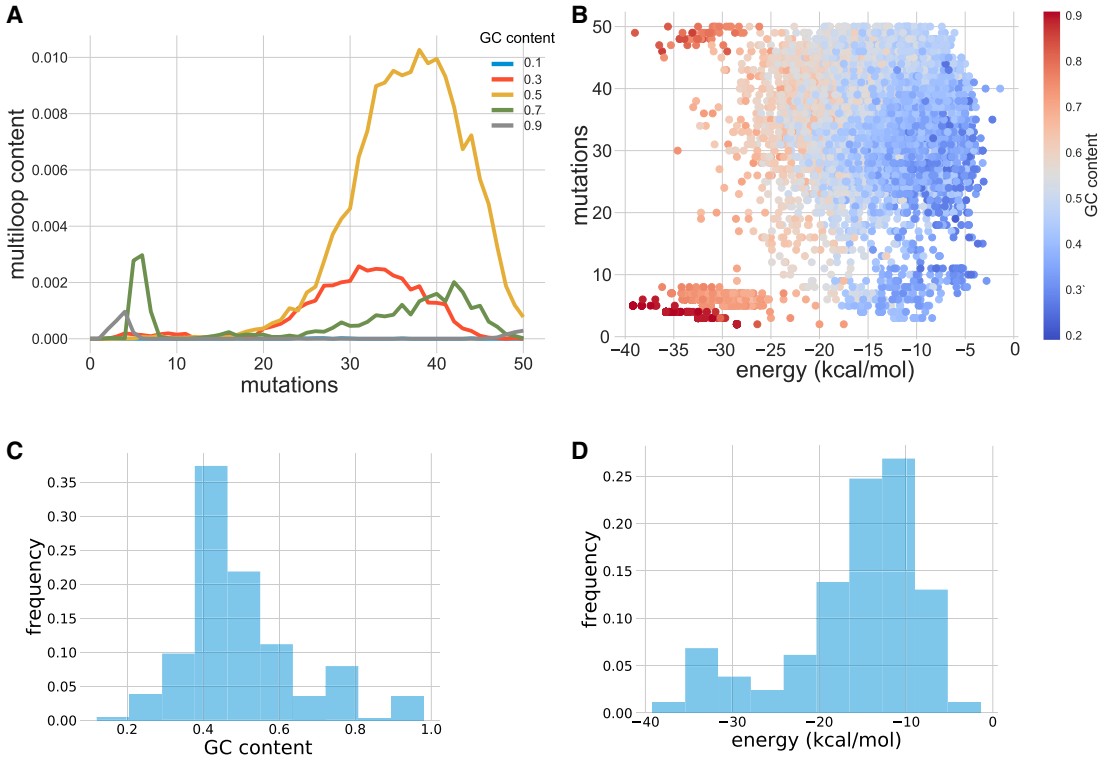

FIGURE 4. Analysis of the distribution of multibranched structures in the RNA mutational landscape sampled by `RNAmutants`. (*A*) The frequency of multibranched structures with respect to the number of mutations from the seed sequence. (*B*) Plot of folding energies and GC contents of each individual multibranched structure. (*C*) Distribution of the GC content of multibranched structures. (*D*) Distribution of folding energies of multibranched structures.

with the energies observed in the Rfam database for structures within this range of GC content values (see Fig. 1B).

Random replication without selection

In the previous section, we observed that the low-energy structural landscape is enriched with ML architectures at specific GC contents. We must next determine under which conditions a replicating population can reach this reservoir. Our first scenario aims to study the behavior and outcome of random replication process *without natural selection*. Here, we consider a simple model in which RNA molecules are duplicated with a small error rate, but preserving the GC content (Tamura 1992). In our simulations, we use an error rate of 0.02 (probability of introducing a point mutation) to allow immediate comparison of the number of elapsed generations, and 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 Materials and Methods). Supplemental Figure S6 shows the results of this calculation for GC content biases varying from 0.1 to 0.9. Noticeably, our data reveals that after a short initialization phase (i.e., after approximately 50 generations), sequences with a GC content of 0.5 have on average slightly more than 35 mutations. This observation is in good adequacy with the peak of multibranched structures identified in Figure 4A,C. This combination of events suggests that the population is randomly exploring the sequence landscape and gets fixed once stable structures are discovered. Indeed, the mutation distance where MLs are observed coincides with the largest mutational neighborhood (see Supplemental Fig. S7), thus where the sequence specificity pressure is minimal. We conclude that a simple undirected replication mechanism could explain an enrichment of RNA populations with stable multibranched structures.

To complete this analysis and assess the different structural compositions between the uniform and low-energy landscapes, we sample sequences at each mutational neighborhood uniformly at random and compute their MFE value and secondary structure. We compare in Figure 5 the average MFE and frequency of MLs in MFE structures between the uniform ("Random") samples and sequences sampled from the `RNAmutants` low-energy ensemble. Importantly, we report separately the statistics for sequences with or without MLs.

Unsurprisingly, the accumulation of mutations does not impact the results in random populations. First, we note that although multibranched structures are relatively

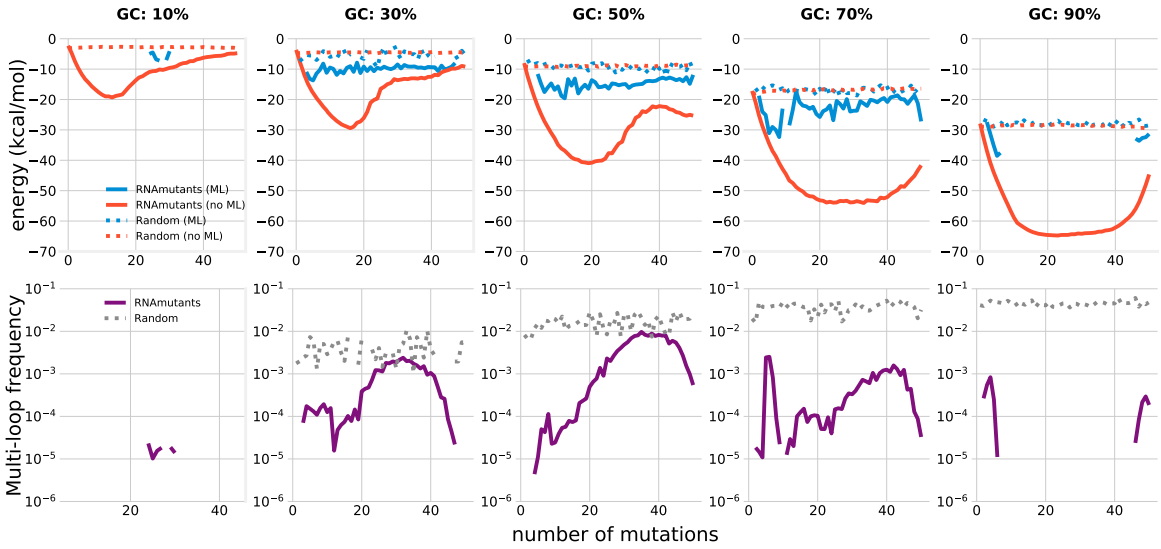

FIGURE 5. Analysis of ML distributions in uniform and low energy populations. First row: average minimum free energies of uniformly sampled mutants ("Random"; dotted lines) and sequences from the low energy ensemble ("`RNAmutants`"; plain lines). Sequences with a MFE structure having a multiloop ("ML"; blue) are separated from the others ("no ML"; orange). Second row: frequency of multibranched structures in the MFE structure of uniformly sampled mutants ("Random"; dotted gray lines) and sequences sampled from the low energy ensemble ("`RNAmutants`"; purple lines)

frequent in random populations (on average between 1% and 10%), they also have higher folding energies than sequences in the low-energy landscape. Higher GC contents tend to increase these trends. This data supports previous observations made on shorter sequences showing that multibranched structures are rare and relatively unstable in random pools of sequences (Stich et al. 2008).

In contrast, `RNAmutants` samples exhibit a different pattern. While multibranched structures are rare in random populations (i.e., no mutations), their frequency increases with the number of mutations (i.e., increased preference toward stable structures). Interestingly, the frequency of stable multibranched structures almost matches 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. 5).

The analysis of folding energies reveals another interesting phenomenon. While the average energies of multibranched structures remain steady at all mutational distances, this is not the case of other structures with simpler architectures (first row of Fig. 5). 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 multibranched structures in the low energy ensemble (i.e., `RNAmutants`) increases due to a better (collective) resilience of this architecture to point-wise mutations and/or more uniform distribution in the sequence landscape. In turn, it increases their density in the large/distant mutational neighborhoods. Moreover, it is worth noting that the values of the folding energies at these GC regimes are also close to those observed in Rfam.

Eventually, we also distinguish a secondary peak of occurrences of multibranched structures in the vicinity of the seeds (i.e., 5–10 mutations) at higher GC regimes (0.7). In 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 multibranched structures in these neighborhoods. We discuss this phenomenon in the section below.

Observation 2

- Multibranched structures are relatively common but unstable in random populations.
- At intermediate GC contents (0.3–0.5), multibranched loops are frequent among the most stable structures.
- The folding energies of stable ML are similar to those observed in databases.

Random replication with selection for stable structures

Our `RNAmutants` and random replication simulations suggest that mutational networks contain reservoirs of

complex structures accessible by undirected replication mechanisms. At this point, our main question is to determine if a natural selection process, independent of a particular target, could help populations reach these regions.

To address this question, we build an evolutionary algorithm (EA) named `mateRNA1`, where the fitness is proportional to the folding energies of the molecules. Intuitively `mateRNA1`, simulates the behavior of a population of RNAs selecting at each generation the most stable sequences regardless of the structure adopted to carry functions (see Materials and Methods). This setting is similar to the energy-based selection described by Fontana and Schuster (1987) on binary sequences. 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 varying mutation rates. Importantly, we also vary the strength of the selective pressure applied on the population (i.e., the β in Equation 1). All simulations were run for 1000 generations.

We show the frequency and folding energy of multibranched structures observed during our simulations with `mateRNA1` (Fig. 6). Here, a higher selective pressure (i.e., larger values of β) and GC content reduces the frequency of MLs but increases the folding energy. Interestingly, a transition occurs when the value of β shifts from 0.01 to 0.05. In contrast, higher GC contents increase both the frequency of MLs and folding energies. Yet, variations of the selective pressure result in more variance of the folding energy.

In our simulations, only high GC contents (≥0.7) produce populations with ML frequencies comparable to those observed in databases. It follows that the `mateRNA1` scenario does not seem fit to explain how to reach multibranched sequences with low to intermediate GC content (0.3 −0.5). However, as noted earlier, there is a second pool of multibranched structures at higher GC content (≥0.7) that are characterized by their proximity to the seed (see Fig.4B; Supplemental Fig. S1) and lower folding energies (see Fig.4D). Under our assumption, the `mateRNA1` scenario appears to be a legitimate candidate to explain how these structures are reached.

We find that `mateRNA1` populations are able to quickly find low energy multibranched structures (less 12 mutations from the initial populations; see Fig. 6), with higher GC content regimes leading to more stable structures. Multibranched structures are mostly found near the seeds (≤20 mutations) and rather in earlier generations.

We also note an interesting dichotomy between populations at 0.7 and 0.9 GC ratios. While the frequency of MLs reached by populations at 0.7 roughly matches the one computed by `RNAmutants` (see Fig. 4A), it is not the case for higher GC contents. Indeed, the simulations find a significant number of MLs not sampled by `RNAmutants`. The folding energies of these structures appear to be similar to the ones found in Rfam (see Fig. 1B).

All together, these observations suggest that selective pressure on stable structures might help in finding stable multibranched architectures at the highest GC contents (i.e., 0.9) that are otherwise eclipsed by the vast number of stable single stem structures. At 0.7 though this selection mechanism could still be used to accelerate and

FIGURE 6. Analysis of ML frequency and energy in `mateRNA1` populations as a function of mutational distance. First row: average frequency of ML structures in populations at each mutational distance. Each line represents a simulation at varying selection strengths (parameter β). Subplot columns correspond to simulations at various GC content biases. Second row: Average population energy at each mutational distance. Plain lines report statistics calculated for multibranched structures only, while dashed lines show the measurements obtained for all structures.

Observation 3

- Multiloops emerge only in populations with the highest GC contents (≥0.7).
- Highest GC contents help to quickly reach MLs in the vicinity of the seeds.
- A selection pressure toward stable structure promotes the discovery of rare stable multibranched structures at 0.9.

promote the discovery of these complex structures uniformly distributed in the sequence landscape.

DISCUSSION

We provided evidence that in the absence of selective pressure, the structure of the mutational landscape could have helped promote the emergence of complex RNA phenotypes. 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 support parsimonious evolutionary scenarios based on undirected molecular replications with occasional mutations. In these simple models, the GC content appears as a key feature in determining the probability of discovering stable multibranched secondary structures. Our study reveals two distinct phenomena. At low to intermediate GC contents (0.3–0.5) the distribution of MLs in a replicating population eliciting sequences with the most stable structures resembles the one observed in RNA databases. This observation suggests an evolutionary scenario in which sequences are replicated without selection until the discovery of complex structures (approximately 30 mutations), at which point selection for stability could begin.

In contrast, at higher GC contents (≥0.7) the presence of multibranched structures appears to require the help of a selective pressure to reach these complex phenotypes. In this work, we simulate replicating RNA populations selecting sequences with the most stable structures at each generation. Then, we show that this mechanism enables a quick discovery of MLs in the vicinity of random sequences at GC content regimes at or above 0.7. This finding is in agreement with previous theoretical studies that showed that neutral networks percolate the whole-sequence landscape (Schuster et al. 1994; Fontana and Schuster 1998). It also suggests a different origin for multibranched structures at high GC content. In such a scenario, a population of molecules would progressively improve the functional efficiency of the molecules by improving their stability.

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 favored the emergence of life. It also offers novel insights into the fundamental properties of the genetic alphabet (Gardner et al. 2003). Incidentally, these observations suggest further investigations into the role of more complex nucleotide distributions (Levin et al. 2012).

It is worth noting that our scenario remains compatible with further selection mechanisms that may come into play once a functional and stable architecture is identified to rapidly improve active sites (Kennedy et al. 2010; Dingle et al. 2015). Eventually, our results could also be used to put in perspective earlier findings suggesting that natural selection is not required to explain pattern composition in rRNAs (Smit et al. 2006).

Our analysis completes recent studies that aimed to characterize fundamental properties of genotype–phenotype maps (Greenbury and Ahnert 2015; Manrubia and Cuesta 2017), and showed that their structure may contribute to the emergence of functional molecules (Dingle et al. 2015). Whereas previous studies focused on characterizing the static genotype–phenotype map of random sequences, we show that the landscape of stable mutants arising from random seeds favors the discovery of complex structures. It also emphasizes the relevance of theoretical models based on a thermodynamical view of prebiotic evolution (Pascal et al. 2013).

The size of the RNA sequences considered in this study has been fixed at 50 nt. This length appears to be the current upper limit for nonenzymatic synthesis (Hill et al. 1993), 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 (Waldispühl et al. 2002). Although, if these variations remain modest, we do not expect any major impact on our conclusions.

The error rates considered in this study were chosen to match the values used in previous related works (e.g., Manrubia and Briones 2006). This choice is also corroborated by recent experiments suggesting that early life scenarios could sustain high error rates (Rajamani et al. 2010). Nevertheless, lower mutation rates would only increase the number of generations needed to reach the asymptotic behavior (see Supplemental Fig. S6), and thus would not affect our results.

Finally, we emphasize that our results do not exclude the use of more advanced evolutionary mechanisms (Szabó et al. 2002; Szathmáry 2006; Briones et al. 2009; Shay et al. 2015). Instead, 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.

MATERIALS AND METHODS

Evolutionary algorithm (`mateRNA1`)

Here, we describe an EA for energy-based selection with GC content bias. The algorithm is implemented in Python and freely available at http://csb.cs.mcgill.ca/maternal.

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 (ω,s) such that s is the MFE structure on sequence ω as computed by the software `RNAfold` version 2.1.9 (Lorenz et al. 2011). Each sequence is formed as a string from the alphabet $B:=\{A,U,C,G\}$. For all experiments, we work with a constant population size of $|P_t| = 1000$ and constant sequence length $\text{len}(s_i) = 50\ \forall s_i \in P_t$. 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.

Initial population

Sequences in the initial population, that is, generation $t = 0$, are generated by sampling sequences of the appropriate length uniformly at random from the alphabet B.

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 ω is defined as the probability $F(\omega,s)$ that ω 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 ω and a given target structure (Stich et al. 2007). However, in our model, this function is proportional to the free energy of the sequence–structure pair, $E(\omega,s)$ as computed by `RNAfold`:

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

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 $N = \sum_{\omega', s' \in S_t} \exp[-\beta E(\omega', s')/RT]$. This normalization enforces that reproduction probability of a sequence–structure pair is weighted against the Boltzmann weight of the entire population. β is the selection strength coefficient. Higher values of β result in stronger selection for stability and vice versa. $R = 0.00019871$ kcal mol^{-1} and $T = 310.15$ K 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 μ known as the mutation rate. μ defines the probability of incorrectly copying a nucleotide and instead randomly sampling one of the other three bases in B.

Controlling population GC content

There are two obstacles to maintaining evolving populations within the desired GC content range of ±0.1. First, an initial population of random sequences sampled uniformly from the full alphabet naturally tends to converge to a GC content of 0.5. To avoid this, we sample from the alphabet with the 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. 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.

`RNAmutants`

The EA implemented in `mateRNA1` 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` (Waldispühl et al. 2008) can sample sequence–structure pairs (ω,s) such that (i) the sequence is a k-mutant from a given seed ω_0—for any k—and (ii) the probability of seeing the pair is proportional to its *fitness* compared to all pairs (ω',s') where ω' is also an k-mutant of ω_0.

In addition, `RNAmutants` provides an unbiased control of the samples GC content allowing direct comparisons with `mateRNA1`.

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 Supplemental Figure S2. Therefore, we replace the sampled structure with the MFE to simplify the study.

For each GC content in {0.1, 0.3, 0.5, 0.7, 0.9} (±0.1) we generated 20 random seeds 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, 12 MB Cache, 95 W) on the Guillimin High Performance Computing Cluster of Calcul Québec. It took over 12,000 CPU hours to complete the sampling.

Sequence–structure pairs weighted sampling

Given a seed sequence ω_0, and a fixed number k of mutations, we denote $\mathbb{S}^k_{\omega_0}$ as the ensemble of all sequence-structure pairs whose hamming distance to ω_0 is k. Similar to the "Fitness function" section, the probability of sampling 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 ω in conformation s is

TABLE 1. Transition probabilities for all pairs of bases where α is the mutation rate and Θ is the desired GC content

	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	$(1-\Theta)\alpha$	$(1-\Theta)\alpha$	$\Theta\alpha$	$1-(1-\alpha(1+\Theta))$

$E(\omega,s)$ then the weight of the pair is

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

whereas before the Boltzmann constant R equalled 0.00019871 kcal mol^{-1} and the temperature T was set at 310.15 K. The normalization factor, or partition function, Z can now be defined as

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

and thus the probability of sampling a pair (ω, s) is

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

By increasing k from 1 to $|\omega_0|(= 50)$ an exploration of the whole-mutational landscape of ω_0 is performed. To compute 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 the complexity of $\mathcal{O}(n^2)$.

Controlling sample 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 (Turner and Mathews 2010) which is used by `RNAmutants`, sampling from any ensemble S will be highly biased toward sequences with high GC content. To get a sample (ω,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 (Waldispühl and Ponty 2011) is applied.

An unbiased sampling of pairs (ω,s) for any given GC target can be obtained by modifying the Boltzmann weights of any element (ω,s) with a term $\boldsymbol{w}^\omega \in [0, 1]$ which depends on the GC content of ω. At its simplest, it can be the proportion of GC in ω. The weight of (ω,s) becomes

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

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

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

To find the weights $\mathbf{w}$ for any target GC an exact solution could be found. In practice, the weight $\boldsymbol{w}^\omega$ is determined as follows, applying the bisection algorithm. Given the number of GC in ω, (resp. number of AU) and two weights $\boldsymbol{w}_{\mathrm{GC}}$ and $\boldsymbol{w}_{\mathrm{AU}}$. We define $\boldsymbol{w}^\omega$ as $\boldsymbol{w}^\omega := \boldsymbol{w}_{\mathrm{GC}}^{|\omega|_{\mathrm{GC}}} \times \boldsymbol{w}_{\mathrm{AU}}^{|\omega|_{\mathrm{AU}}}$. At the first iteration, a thousand sequences are sampled with $\boldsymbol{w}_{\mathrm{GC}}$ and $\boldsymbol{w}_{\mathrm{AU}}$ both at 0.5. If the average GC content on the sampled sequences is too high, the value of $\boldsymbol{w}_{\mathrm{GC}}$ is decreased by half and $\boldsymbol{w}_{\mathrm{AU}}$ increased accordingly. If the average GC content is too low, the opposite is done. Each time, the sequences with the desired GC content are kept. This process is repeated until the desired amount of sequences with the required GC content is produced. In practice, after a few iterations almost all sampled sequences contain the target GC content and as shown this sampling is unbiased (Waldispühl and Ponty 2011). An interesting observation is that the same method can be applied to preferentially sample sequences with any other desired feature.

Sequence divergence in a random replication model

We estimate the expected number of mutations in randomly replicated sequences (see the "Random replication without selection" section) using the transition matrix defined by Tamura (1992). We use a mutation rate $\alpha = 0.02$ mirroring the mutation rate used in `mateRNA1`, 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 in Table 1.

This matrix gives us the transition rate from one generation to the next one. To obtain the mutation probabilities at the kth generation, we calculate the kth 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 kth generation.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

C.G.O. is supported by a Fonds de Recherche Nature et Technologie Quebec (FRQNT) Doctoral Fellowship. V.R. is supported by a Fonds de Recherche Nature et Technologie Quebec (FRQNT) and Azrieli Postdoctoral Fellowships. J.W. is supported by a NSERC Discovery grant (RGPIN-2015-03786), NSERC Accelerator Award (RGPAS 477873-15), and FRQ-NT INRIA Associated Teams (211485).

Author contributions: C.G.O., V.R., and J.W. designed the research, analyzed the results, and wrote the manuscript. C.O. and V.R. conducted the computational experiments.

Received January 15, 2019; accepted August 19, 2019.

REFERENCES

Adamala K, Szostak JW. 2013. Nonenzymatic template-directed RNA synthesis inside model protocells. *Science* **342:** 1098–1100. doi:10.1126/science.1241888

Aguirre J, Buldú JM, Stich M, Manrubia SC. 2011. Topological structure of the space of phenotypes: the case of RNA neutral networks. *PLoS One* **6:** e26324. doi:10.1371/journal.pone.0026324

Ancel LW, Fontana W. 2000. Plasticity, evolvability, and modularity in RNA. *J Exp Zool* **288:** 242–283. doi:10.1002/1097-010X(20001015)288:3<242::AID-JEZ5>3.0.CO;2-O

Bartel DP, Szostak JW. 1993. Isolation of new ribozymes from a large pool of random sequences. *Science* **261:** 1411–1418. doi:10.1126/science.7690155

Beaudry AA, Joyce GF. 1992. Directed evolution of an RNA enzyme. *Science* **257:** 635–641. doi:10.1126/science.1496376

Becker S, Thoma I, Deutsch A, Gehrke T, Mayer P, Zipse H, Carell T. 2016. A high-yielding, strictly regioselective prebiotic purine nucleoside formation pathway. *Science* **352:** 833–836. doi:10.1126/science.aad2808

Briones C, Stich M, Manrubia SC. 2009. The dawn of the RNA World: toward functional complexity through ligation of random RNA oligomers. *RNA* **15:** 743–749. doi:10.1261/rna.1488609

Chen IA, Salehi-Ashtiani K, Szostak JW. 2005. RNA catalysis in model protocell vesicles. *J Am Chem Soc* **127:** 13213–13219. doi:10.1021/ja051784p

Cowperthwaite MC, Economo EP, Harcombe WR, Miller EL, Meyers LA. 2008. The ascent of the abundant: how mutational networks constrain evolution. *PLoS Comput Biol* **4:** e1000110. doi:10.1371/journal.pcbi.1000110

Darty K, Denise A, Ponty Y. 2009. VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25:** 1974–1975. doi:10.1093/bioinformatics/btp250

Dingle K, Schaper S, Louis AA. 2015. The structure of the genotype–phenotype map strongly constrains the evolution of non-coding RNA. *Interface Focus* **5:** 20150053. doi:10.1098/rsfs.2015.0053

Eddy SR. 2001. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet* **2:** 919–929. doi:10.1038/35103511

Ferré-D'Amaré AR, Scott WG. 2010. Small self-cleaving ribozymes. *Cold Spring Harb Perspect Biol* **2:** a003574. doi: 10.1101/cshperspect.a003574

Fontana W, Schuster P. 1987. A computer model of evolutionary optimization. *Biophys Chem* **26:** 123–147. doi:10.1016/0301-4622(87)80017-0

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

Fontana W, Griesmacher T, Schnabl W, Stadler PF, Schuster P. 1991. Statistics of landscapes based on free energies, replication and degradation rate constants of RNA secondary structures. *Monatshefte für Chemie/Chem Mon* **122:** 795–819. doi:10.1007/BF00815919

Gardner PP, Holland BR, Moulton V, Hendy M, Penny D. 2003. Optimal alphabets for an RNA world. *Proc Biol Sci* **270:** 1177–1182. doi:10.1098/rspb.2003.2355

Gilbert W. 1986. Origin of life: the RNA world. *Nature* **319:** 618. doi:10.1038/319618a0

Greenbury SF, Ahnert SE. 2015. The organization of biological sequences into constrained and unconstrained parts determines fundamental properties of genotype–phenotype maps. *J R Soc Interface* **12:** 20150724. doi:10.1098/rsif.2015.0724

GrunerW, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Stadler PF, Schuster P. 1996. Analysis of RNA sequence structure maps by exhaustive enumeration I. Neutral networks. *Monatsh Chem* **127:** 355–374. doi:10.1007/BF00810881

Hayden EJ, Lehman N. 2006. Self-assembly of a group I intron from inactive oligonucleotide fragments. *Chem Biol* **13:** 909–918. doi:10.1016/j.chembiol.2006.06.014

Higgs PG, Lehman N. 2015. The RNA World: molecular cooperation at the origins of life. *Nat Rev Genet* **16:** 7–17. doi:10.1038/nrg3841

Hill AR Jr, Orgel LE, Wu T. 1993. The limits of template-directed synthesis with nucleoside-5′-phosphoro(2-methyl)imidazolides. *Orig Life Evol Biosph* **23:** 285–290. doi:10.1007/BF01582078

Horning DP, Joyce GF. 2016. Amplification of RNA by an RNA polymerase ribozyme. *Proc Natl Acad Sci* **113:** 9786–9791. doi:10.1073/pnas.1610103113

Ivica NA, Obermayer B, Campbell GW, Rajamani S, Gerland U, Chen IA. 2013. The paradox of dual roles in the RNA world: resolving the conflict between stable folding and templating ability. *J Mol Evol* **77:** 55–63. doi:10.1007/s00239-013-9584-x

Kennedy R, Lladser ME, Wu Z, Zhang C, Yarus M, De Sterck H, Knight R. 2010. Natural and artificial RNAs occupy the same restricted region of sequence space. *RNA* **16:** 280–289. doi:10.1261/rna.1923210

Kupczok A, Dittrich P. 2006. Determinants of simulated RNA evolution. *J Theor Biol* **238:** 726–735. doi:10.1016/j.jtbi.2005.06.019

Leu K, Obermayer B, Rajamani S, Gerland U, Chen IA. 2011. The prebiotic evolutionary advantage of transferring genetic information from RNA to DNA. *Nucleic Acids Res* **39:** 8135–8147. doi:10.1093/nar/gkr525

Levin A, Lis M, Ponty Y, O'Donnell CW, Devadas S, Berger B, Waldispühl J. 2012. A global sampling approach to designing and reengineering RNA secondary structures. *Nucleic Acids Res* **40:** 10041–10052. doi:10.1093/nar/gks768

Levy M, Ellington AD. 2001. The descent of polymerization. *Nat Struct Biol* **8:** 580–582. doi:10.1038/89601

Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA package 2.0. *Algorithms Mol Biol* **6:** 26. doi:10.1186/1748-7188-6-26

ManrubiaSC, Briones C. 2006. Modular evolution and increase of functional complexity in replicating RNA molecules. *RNA* **13:** 97–107. doi:10.1261/rna.203006

Manrubia S, Cuesta JA. 2017. Distribution of genotype network sizes in sequence-to-structure genotype–phenotype maps. *J R Soc Interface* **14:** 20160976. doi:10.1098/rsif.2016.0976

Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, et al. 2015. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res* **43** (Database issue)**:** D130–D137. doi:10.1093/nar/gku1063

O'Flaherty DK, Kamat NP, Mirza FN, Li L, Prywes N, Szostak JW. 2018. Copying of mixed-sequence RNA templates inside model protocells. *J Am Chem Soc* **140:** 5171–5178. doi:10.1021/jacs.8b00639

O'Flaherty DK, Zhou L, Szostak JW. 2019. Nonenzymatic template-directed synthesis of mixed-sequence 3′-NP-DNA up to 25 nucleotides long inside model protocells. *J Am Chem Soc* **141:** 10481–10488. doi:10.1021/jacs.9b04858

Orgel LE. 2004. Prebiotic chemistry and the origin of the RNA world. *Crit Rev Biochem Mol Biol* **39:** 99–123. doi:10.1080/10409230490460765

Pascal R, Pross A, Sutherland JD. 2013. Towards an evolutionary theory of the origin of life based on kinetics and thermodynamics. *Open Biol* **3:** 130156. doi:10.1098/rsob.130156

Paul N, Joyce GF. 2002. A self-replicating ligase ribozyme. *Proc Natl Acad Sci* **99:** 12733–12740. doi:10.1073/pnas.202471099

Pearce BKD, Pudritz RE, Semenov DA, Henning TK. 2017. Origin of the RNA world: the fate of nucleobases in warm little ponds. *Proc Natl Acad Sci* **114:** 11327–11332. doi:10.1073/pnas.1710339114

Penny D, Poole A. 1999. The nature of the last universal common ancestor. *Curr Opin Gen Dev* **9:** 672–677. doi:10.1016/S0959-437X(99)00020-9

Powner MW, Gerland B, Sutherland JD. 2009. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* **459:** 239–242. doi:10.1038/nature08013

Pressman A, Moretti JE, Campbell GW, Müller UF, Chen IA. 2017. Analysis of in vitro evolution reveals the underlying distribution of catalytic activity among random sequences. *Nucleic Acids Res* **45:** 8167–8179. doi:10.1093/nar/gkx540

Rajamani S, Ichida JK, Antal T, Treco DA, Leu K, Nowak MA, Szostak JW, Chen IA. 2010. Effect of stalling after mismatches on the error catastrophe in nonenzymatic nucleic acid replication. *J Am Chem Soc* **132:** 5880–5885. doi:10.1021/ja100780p

Reidys C, Stadler PF, Schuster P. 1997. Generic properties of combinatory maps: neutral networks of RNA secondary structures. *Bull Math Biol* **59:** 339–397. doi:10.1007/BF02462007

Ritson D, Sutherland JD. 2012. Prebiotic synthesis of simple sugars by photoredox systems chemistry. *Nat Chem* **4:** 895–899. doi:10.1038/nchem.1467

Robertson MP, Joyce GF. 2012. The origins of the RNA world. *Cold Spring Harb Perspect Biol* **4:** a003608. doi:10.1101/cshperspect.a003608

Robertson MP, Joyce GF. 2014. Highly efficient self-replicating RNA enzymes. *Chem Biol* **21:** 238–245. doi:10.1016/j.chembiol.2013.12.004

Salehi-Ashtiani K, Szostak JW. 2001. In vitro evolution suggests multiple origins for the hammerhead ribozyme. *Nature* **414:** 82–84. doi:10.1038/35102081

Schultes EA, Spasic A, Mohanty U, Bartel DP. 2005. Compact and ordered collapse of randomly generated RNA sequences. *Nat Struct Mol Biol* **12:** 1130–1136. doi:10.1038/nsmb1014

Schuster P. 2001. Evolution in silico and in vitro: the RNA model. *Biol Chem* **382:** 1301–1314. doi:10.1515/BC.2001.162

Schuster P. 2006. Prediction of RNA secondary structures: from theory to models and real molecules. *Rep Prog Phys* **69:** 1419–1477. doi:10.1088/0034-4885/69/5/R04

Schuster P, Fontana W. 1999. Chance and necessity in evolution: lessons from RNA. *Physica D* **133:** 427–452. doi:10.1016/S0167-2789(99)00076-7

Schuster P, Stadler PF. 1994. Landscapes: complex optimization problems and biopolymer structures. *Comput Chem* **18:** 295–324. doi:10.1016/0097-8485(94)85025-9

Schuster P, Fontana W, Stadler PF, Hofacker IL. 1994. From sequences to shapes and back: a case study in RNA secondary structures. *Proc Biol Sci* **255:** 279–284. doi:10.1098/rspb.1994.0040

Shapiro R. 2007. A simpler origin for life. *Sci Am* **296:** 46–53. doi:10.1038/scientificamerican0607-46

Shay JA, Huynh C, Higgs PG. 2015. The origin and spread of a cooperative replicase in a prebiotic chemical system. *J Theor Biol* **364:** 249–259. doi:10.1016/j.jtbi.2014.09.019

Smit S, Yarus M, Knight R. 2006. Natural selection is not required to explain universal compositional patterns in rRNA secondary structure categories. *RNA* **12:** 1–14. doi:10.1261/rna.2183806

Stich M, Briones C, Manrubia SC. 2007. Collective properties of evolving molecular quasispecies. *BMC Evol Biol* **7:** 110. doi:10.1186/1471-2148-7-110

Stich M, Briones C, Manrubia SC. 2008. On the structural repertoire of pools of short, random RNA sequences. *J Theor Biol* **252:** 750–763. doi:10.1016/j.jtbi.2008.02.018

Stich M, Manrubia SC, La E. 2010. Variable mutation rates as an adaptive strategy in replicator populations. *PLoS One* **5:** e11186. doi:10.1371/journal.pone.0011186

Szabó P, Scheuring I, Czárán T, Szathmáry E. 2002. In silico simulations reveal that replicators with limited dispersal evolve towards higher efficiency and fidelity. *Nature* **420:** 340–343. doi:10.1038/nature01187

Szathmáry E. 2006. The origin of replicators and reproducers. *Philos Trans R Soc Lond B Biol Sci* **361:** 1761–1776. doi:10.1098/rstb.2006.1912

Szostak JW. 2012. The eightfold path to non-enzymatic RNA replication. *J Syst Chem* **3:** 2. doi:10.1186/1759-2208-3-2

Tamura K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition–transversion and G+C-content biases. *Mol Biol Evol* **9:** 678–687.

Turk RM, Chumachenko NV, Yarus M. 2010. Multiple translational products from a five-nucleotide ribozyme. *Proc Natl Acad Sci* **107:** 4585–4589. doi:10.1073/pnas.0912895107

TurnerDH, Mathews DH. 2010. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res* **38(**Suppl 1**):** D280–D282. doi:10.1093/nar/gkp892

Vaidya N, Manapat ML, Chen IA, Xulvi-Brunet R, Hayden EJ, Lehman N. 2012. Spontaneous network formation among cooperative RNA replicators. *Nature* **491:** 72–77. doi:10.1038/nature11549

van Nimwegen E, Crutchfield JP, Huynen M. 1999. Neutral evolution of mutational robustness. *Proc Natl Acad Sci* **96:** 9716–9720. doi:10.1073/pnas.96.17.9716

Waldispühl J, Behzadi B, Steyaert J-M. 2002. An approximate matching algorithm for finding (sub-) optimal sequences in S-attributed grammars. *Bioinformatics* **18**(Suppl 2)**:** S250–S259. doi:10.1093/bioinformatics/18.suppl_2.S250

Waldispühl J, Devadas S, Berger B, Clote P. 2008. Efficient algorithms for probing the RNA mutation landscape. *PLoS Comput Biol* **4:** e1000124. doi:10.1371/journal.pcbi.1000124

Waldispühl J, Ponty Y. 2011. An unbiased adaptive sampling algorithm for the exploration of RNA mutational landscapes under evolutionary pressure. *J Comput Biol* **18:** 1465–1479. doi:10.1089/cmb.2011.0181

Wilke CO. 2001. Selection for fitness versus selection for robustness in RNA secondary structure folding. *Evolution* **55:** 2412–2420. doi:10.1111/j.0014-3820.2001.tb00756.x