Author's Accepted Manuscript

www.elsevier.com/locate/yjtbi

Motif frequency and evolutionary search times in RNA populations

Michael Stich, Susanna C. Manrubia

PII: S0022-5193(11)00150-0
DOI: doi:10.1016/j.jtbi.2011.03.010
Reference: YJTBI 6403

To appear in: *Journal of Theoretical Biology*

Received date: 11 August 2010
Revised date: 26 January 2011
Accepted date: 10 March 2011

Cite this article as: Michael Stich and Susanna C. Manrubia, Motif frequency and evolutionary search times in RNA populations, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2011.03.010

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Motif frequency and evolutionary search times in RNA populations

Michael Stich*, Susanna C. Manrubia

Centro de Astrobiología (CSIC-INTA), Ctra de Ajalvir km 4, 28850 Torrejón de Ardoz (Madrid), Spain.

Abstract

RNA molecules, through their dual identity as sequence and structure, are an appropriate experimental and theoretical model to study the genotype–phenotype map and evolutionary processes taking place in simple replicator populations. In this computational study, we relate properties of the sequence–structure map, in particular the abundance of a given secondary structure in a random pool, with the number of replicative events that an initially random population of sequences needs to find that structure through mutation and selection. For common structures, this search process turns out to be much faster than for rare structures. Furthermore, search and fixation processes are more efficient in a wider range of mutation rates for common structures, thus indicating that evolvability of RNA populations is not simply determined by abundance. We also find significant differences in the search and fixation processes for structures of same abundance, and relate them with the number of base pairs forming the structure. Moreover, the influence of the nucleotide content of the RNA sequences on the search process is studied. Our results advance in the understanding of the distribution and attainability of RNA secondary structures. They hint at the fact that, beyond sequence length and sequence-to-function redundancy, the mutation rate that permits localization and fixation of a given phenotype strongly depends on its relative abundance and global, in general nonuniform, distribution in sequence space.

Keywords: Genotype–phenotype map, RNA secondary structure, RNA world

Accepted manuscript

1. Introduction

RNA molecules are a very well-suited model for studying evolving populations of replicators because they incorporate, in a single molecular entity, both genotype and phenotype. While errors in the replication process introduce mutations in the RNA sequence (genotype), selection acts upon the function (phe-

*Corresponding author. Tel.: +34 91 520 6409; Fax: +34 91 520 1074
Email address: stichm@inta.com (Michael Stich)

ACCEPTED MANUSCRIPT

notype) of the molecule. Since the biochemical function of RNA to large extent is given by its three-dimensional conformation, the genotype-to-phenotype map of RNA can be splitted conceptually into a map from sequence to structure and a map from structure to function.

At the heart of any evolutionary RNA model lies the mapping from an RNA sequence to a structure, the folding process. While the folding process of RNA sequence is complex and finally yields three-dimensional tertiary structures, the planar secondary structure is a folding intermediate and represents a building block of the tertiary structure. Therefore, it is often justified to use secondary structure as a proxy for function. Using secondary structure as approximation of tertiary structure is particularily well justified for short sequences, since the probability of displaying tertiary contacts increases with the sequence length. Nevertheless, even for a molecule as large as the complete HIV-1 genome, secondary structure has been determined and functional RNA motifs identified (Watts et al., 2009). To clarify the intimate relation of RNA structure to function is a very active field of research and fitness landscapes of real RNA structures can be constructed empirically (Held et al., 2003; Pitt and Ferré-D'Amaré, 2010). The RNA folding process depends on many parameters like temperature or ionic conditions, and there exists a whole ensemble of accessible structures for a given sequence. Nevertheless, in first approximation the structure formed most likely is the minimum free energy structure, used in this article. The prediction of RNA secondary structures is reviewed in Schuster (2006).

Many other aspects of the RNA sequence–structure map have been studied over the decades (Fontana et al., 1993; Schuster et al., 1994; Schuster, 2003), building on the fact that there are many more different sequences than structures. One important observation is that there are few common structures, with many sequences as preimages, and many rare structures, with only few sequences adopting those structures, this revealing an uneven fragmentation of the space of sequences into secondary structures. Furthermore, basic properties of RNA secondary structures are known, e.g., as how the mean number of loops or stems, and their sizes, vary as a function of the length of the molecule (Fontana et al., 1993). Analytical approaches often do not consider energetic aspects of the folded state, but if included can yield useful bounds to the statistical behaviour (Hofacker et al., 1998; Clote et al., 2009). Only for short molecules, where extensive folding of the sequence space can be performed, a complete picture of these quantities can be given (Grüner et al., 1996a,b). For larger molecules, the analysis has to be restricted to subsets of the sequence space.

The size of the pool of considered sequences is of crucial importance when theoretical or computational results are compared to experiments. In SELEX experiments, aimed to produce aptamers with a certain function and hence structure, both the size of the pool and the length of the molecules are important issues in determining the outcome of the selection process, as experimental and theoretical work has shown (Bartel and Szostak, 1993; Sabeti et al., 1997; Gevertz et al., 2005). Furthermore, the structural repertoire of the pool strongly depends on its nucleotide composition (Knight et al., 2005; Kim et al.,

Accepted manuscript

ACCEPTED MANUSCRIPT

2007). Through an extensive computational study of 23 RNA motifs, it was shown recently that natural and artificial RNAs occupy the same region of sequence space (Kennedy et al., 2010). Results from different experimental approaches show that evolved aptamers (Carothers et al., 2004) or small catalytic RNAs (Puerta-Fernández et al., 2003) tend to have simple topologies. Simple structures like stem-loops and hairpins are also found to be more abundant and stable in non-coding regions of prokaryotic genomes than expected by chance (Petrillo et al., 2006). Similarly, a large fraction of non-coding RNAs in long vertebrate genomes fold into simple stem-loop and hairpin structures (Pedersen et al., 2006). In a recent article (Stich et al., 2008), we studied the structural repertoire of an ensemble of 10^8 random RNA sequences. Most abundant structural motifs are topologically simple and occur frequently in natural RNAs (Gan et al., 2003; Cowperthwaite et al., 2008). They could have also played a relevant role in prebiotic scenarios where random polymerization was likely, by constituting simple building blocks able to combine into more complex structures (Briones et al., 2009).

In the context of using RNA as model for evolution, and the aforementioned duality of RNA sequence as bearer of the genotype and RNA secondary structure as phenotype, it is straightforward to implement mutation of bases as a way of introducing variability in the genotype and selection as an operation acting on the secondary structure. While in nature selection is not directed towards a unique, optimal phenotype, and only responds to relative advantages between peers at a fixed time, in simulations we can fix a specific secondary structure as target of selection. Such kind of evolutionary model has been introduced by Fontana and Schuster (1987) and used thereafter for a wide range of purposes and in many variants (e.g., (Huynen et al., 1993, 1996; Ancel and Fontana, 2000; Stich et al., 2007)). As result of such type of evolutionary models, sequences that fold into a given target structure are found. In principle, this can be perceived as a particular solution of the fundamental design problem which consists of selecting the RNA sequence that will adopt a particular secondary structure. For a comparison of computational methods of sequence design, see Dirks et al. (2004).

In this contribution, we investigate how evolutionary processes depend on the choice of the target structure. In a broader context, the target structure is a simplified representation of a functional phenotype, such that with due caution our results can be applicable to more general situations. We work with short sequences for different reasons. First, they are relevant in early chemical (prebiotic) evolution, where high mutation rates limited the length of molecules able to carry genomic information. This is one of our scenarios of interest. Second, the prediction of secondary structures is much more reliable for short sequences, since the probability of tertiary interactions is the lower the shorter the sequence. Last, but not least, by using short sequences, computations are efficient enough to obtain well defined statistical measures of the collective behaviour of the populations studied. Though a recently published program permits to map RNA sequence to structure in polynomial time (Waldispühl et al., 2008) – and hence allows for longer molecules – the necessity to access the minimum free energy

Accepted manuscript

ACCEPTED MANUSCRIPT

structure makes it impossible to use this algorithm. In our analysis, we consider two characteristic temporal quantities. The first is the *search time*, the number of replicative events that an initially random population needs to find the target structure. The second is the *search-plus-fixation time*, the number of generations a population needs to fix the target structure permanently within the population. In this work, we try to shed light on the relation between the range of secondary structures present in a pool of RNA molecules and the time it takes for an evolving population to find a desired secondary structure, i.e., function. By systematically choosing many different target structures that vary in abundance within a random pool, in complexity of their structure, in the number of base pairs, and in their nucleotide composition, the quantities defined above elucidate the influence of the sequence–structure map on the evolutionary dynamics. Besides the already mentioned time scales, we also discuss the range of mutation rates where evolutionary success, i.e., the fixation of the target structure is possible, and how this range depends on the abundance and structural complexity of the target phenotype.

The article is organized in the following way. First, we describe the structural repertoire of a random pool of RNA molecules. Then, we introduce an evolutionary algorithm, where a secondary structure is chosen as target for evolution. Using many different structures, we find that the search for common structures is much more efficient than for rare ones, probably an expected result. However, we show in the subsequent sections that besides abundance other parameters strongly affect the evolvability of RNA secondary structures. We explore how the search process depends on mutation rate, population size, fitness landscape and selection algorithm, number of total base pairs, hairpin loop size, and nucleotide content. The article is closed by a discussion of the results.

2. Structural repertoire of random RNA pools

In this section we review some important properties of pools of random RNA sequences, with special emphasis on their structural diversity.

As a sequence folds, it rapidly attains the secondary structure through formation of base pairs. Then, on a longer time scale, long-range interactions between different parts of the secondary structure may take place and yield a tertiary structure. However, since the secondary structure is a biochemically relevant intermediate and since a large part of the total folding energy is found in the secondary structure, it is regarded as good proxy for biochemical function, especially for short molecules (Schuster, 2003). We are interested in mapping a given sequence to exactly one structure. To this end, we use the minimum free energy structure (see Methods) and do not consider suboptimal foldings, partition functions, or kinetic foldings.

Even in its simplest model, the mapping from sequence to secondary structure is many-to-one and already complex from a theoretical point of view. A neutral network is defined as the subset of all sequences sharing their minimum free energy secondary structure. Usually it is assumed that two sequences in a neutral network are mutually accessible if it is possible to attain one from the

Accepted manuscript

ACCEPTED MANUSCRIPT

other by means of single point mutations in the sequence. If there are sequences in a neutral network which cannot be connected in this way, the neutral network is then formed by disconnected groups of sequences (Grüner et al., 1996b). Following (Fontana et al., 1993; Schuster et al., 1994; Schuster, 2003), we review some basic properties of the RNA landscapes: (a) There are many less structures than sequences and the set of structures can be divided into common and rare structures: While there are 4^n sequences of length n (fully spanning the sequence space), the number S_n of different structures (the structure space) is much smaller. Based on theoretical studies (Waterman, 1978), the expression $S_n \approx 0.7131 \times n^{-3/2}(2.2888)^n$ has been obtained as an upper bound to the number of structures with at least three unpaired nucleotides in a terminal loop and where a single base pair can form a stack (Grüner et al., 1996a; Hofacker et al., 1998). One of the fundamental results in this context is that the distribution of RNA structures within pools of random sequences is very biased, as theoretical studies and observation of natural secondary structures demonstrate (Fontana et al., 1993; Schuster et al., 1994): common structures are typically many orders of magnitude more frequent than rare structures, and there are much less common structures than rare ones (Schuster et al., 1994; Grüner et al., 1996a; Joyce, 2004). (b) Sequences that fold into common structures seem to be randomly distributed in sequence space. This hypothesis results from the observation that a common structure can be found within a relatively small neighborhood of any sequence, a feature called *shape space covering*. (c) Neutral networks of common structures are connected if the sequences are long enough. Furthermore, statistical properties of the folded structures, such as number and sizes of stacks and loops as function of the sequence length have been computed (Fontana et al., 1993). Applying different folding algorithms, it has been shown that structure statistics is relatively insensitive to the precise folding algorithm used (Tacker et al., 1996).

In a previous work (Stich et al., 2008), we described the results of the folding of $M = 10^8$ random RNA sequences of length $n = 35$ nucleotides (nt). As secondary structure of each molecule, we took the minimum free energy structure as given by the `fold()` routine from the Vienna RNA Package (Hofacker et al., 1994). These sequences fold into roughly 5 million different secondary structures. While some of the most frequent structures are produced by tens of thousands of different sequences, the majority of structures are rare, and just represented by 1 or 2 sequences. In the following, we will use the absolute frequency f of a structure in the random pool, and its relative representation $r = f/M$. While common structures are easily obtained, even in small populations, and do not depend strongly on the mean nucleotide composition of the pool (since there exist many sequences compatible with common structures), sequences folding into rare structures often need to be designed, for instance by means of inverse folding algorithms (Schuster et al., 1994; Hofacker et al., 1994).

Our classification of secondary structures relies on the number of basic structural elements forming secondary structures, i.e, hairpin loops, stacks, bulges and interior loops, and multiloops. We found that only 21 *structure families* are enough to cover all 5 million different structures found in the random pool. In

Accepted manuscript

ACCEPTED MANUSCRIPT

increasing order of complexity, we have stem-loops (SL), simple hairpins (HP), hairpins with more structural elements, denoted as HPx, with x standing for the total number of bulges and/or interior loops, double stem-loops (DSL), also with increasing number of bulges/interior loops (DSLx), hammerhead (HH and HHx) structures, triple stem-loops (TSL and TSLx), and others. Using this classification, we were able to determine the distribution of structure families as function of structure rank. The most frequent structures belong to the SL family, followed by the HP, HP2 and HP3 families. In Table 1, we give representative examples of the most relevant structure families – usually they are the most abundant structures as found in (Stich et al., 2008) –, together with their abundance f. These structures are used as target structures in the following sections.

*** Table 1 NEAR HERE ***

3. Evolutionary algorithm

The folding of random sequences yields a static picture of the sequence–structure map. For reasonably long molecules, locating a precise structure through this procedure is very unlikely in case it represents a very rare phenotype. The search can be efficiently performed through evolutionary dynamics, a process that becomes possible only once replication within a population has arisen, that is, once evolution through Darwinian selection is triggered.

Our model consists of a population of N replicating RNA sequences, each of length $n = 35$ nt. At the beginning of the simulation, every molecule of the population is initialized with a random sequence. As a molecule replicates, each nucleotide has a probability (mutation rate) μ to be randomly replaced by another (or the same) type of nucleotide.

At each generation, the sequences are folded into secondary structures as described (see Methods). We define a target secondary structure (or motif) which represents in a simple way optimal performance in the given environment. Every folded structure i is compared with the target structure by means of the base-pair distance d_i, defined as the number of base pairs that have to be opened and closed to transform the given structure into the target structure (Hofacker et al., 1994). The closer a secondary structure is to the target structure, the higher is the probability $p(d_i)$ that the corresponding sequence i replicates. This probability is given by

$$p(d_i) = Z^{-1} \exp(-\beta d_i). \tag{1}$$

The parameter β denotes the strength of selective pressure and is here chosen as $\beta = 2/n$. The distance d scales with the length of the molecule. To avoid a simple dependence on n, we rescale β by the length of the molecule. The normalization factor is $Z = \sum_{i=1}^{N} \exp(-\beta d_i)$. Generations in our simulations are non-overlapping and the offspring generation is calculated according to Wright-Fisher sampling at each time step.

Starting with a population of random sequences, the population first evolves through a search regime. Usually, the average distance of the population to the

ACCEPTED MANUSCRIPT

target structure, $d = \sum_i^N d_i/N$, decreases during this phase. Then, at generation g, for the first time a molecule folds into the target structure. However, due to the stochastic nature of mutation, the population may lose again the target structure. Nevertheless, if the mutation rate is not too large (below the fixation threshold), the average number of correctly folded molecules increases within the population and the target structure gets fixed at generation g_{F}. Eventually, the population reaches an asymptotic regime characterized by statistically constant values for d and ρ, the fraction of molecules folding into the target structure. In absence of an analytic theory for the system, we determine the fixation threshold as the value μ_{F} at which the curve $g_{\mathrm{F}}(\mu)$ diverges. In previous work, we have described several aspects of the evolutionary dynamics of populations subjected to such type of evolutionary algorithm (Stich et al., 2007, 2010a,b).

4. Search times as function of the mutation rate

Based on the classification of the structure families (Stich et al., 2008), we can check how the shape of the secondary structure and its frequency in a random pool affect the search time needed to find it. As structures, we use the most abundant structure (MAS) of 16 structure families plus a hairpin and hammerhead structure previously used in (Stich et al., 2007). The 16 structure families represent all sequences of the pool ($M = 10^8$), with exception of 521 sequences that belong to 5 rare structure families (Stich et al., 2008). The chosen structures are given in Table 1.

*** Figure 1 NEAR HERE ***

In Fig. 1(a), we show for the 18 structures how the search time varies as function of the mutation rate. We observe strong differences among the structures. While some structures are found very easily, almost independently of the mutation rate, others are found only in a limited range of μ and only after a long search process. In general, a search time curve is U-shaped, with an increase of the search time for low mutation rates – only little diversity is introduced, and for large mutation rates – too many deleterious mutations interrupt the path from random to suboptimal structures and further to the target structure.

The curve for the MAS of the SL family, the most abundant structure found in the random pool lies at $g \approx 1$ for all μ. This structure is so frequent that even the starting pool already contains on average 6 sequences that fold into that structure. Also the MAS of the HP family is either already present or found after very few replicative events. Therefore, these curves do not bend up for any μ.

The two next-higher curves correspond to the MAS of the HP2 and DSL families. We see that for small values of μ, the values of g increase as expected. However, for large μ, the curves still do not bend up. As μ increases the selection and search process becomes less efficient and in the limit $\mu \to 1$ at each generation a new random population is produced. There, a common structure with relative representation r can be expected on average to be found after approximately $g \approx 1/(rN)$ generations. In this case, we calculate $g \approx 10$ which

ACCEPTED MANUSCRIPT

agrees well with the values obtained numerically. This means that common structures are found easily, almost independently of the mutation rate.

The next-higher curves correspond to families HP3 and DSL2. Both structures are found equally fast, approximately after 14 generations over large ranges of μ. Both structures have not only a comparable relative representation in a random pool, but also coincide in the total number of base pairs (like also the MAS of HP2 and DSL families). This motivates us to study further below not only the effect that the structure frequency has on the search process (Sec. 6), but also the influence of the number of base pairs for structures with equal abundance (Sec. 7).

For less frequent structures, the curves shift upwards and bend into the U-shape. In general, MAS of higher-order families are found more slowly than MAS of simple structure families. Also the interval of μ where the search is effective becomes narrower as the curve move upwards. It is important to note that due to these effects, a frequent structure has a two-fold advantage in a search process compared to a rare one: lower g for all μ, and larger interval of μ where g is close to its minimum value.

This brings us to the question of whether the population for the considered mutation rate actually fixes the target structure. In Fig. 1(b), we show for five (common and rare) structures both search and search-plus-fixation (short: fixation) times. We determine the fixation time as the generation g_{F} after which at least one molecule of the population folds into the target structure uninterruptedly for 500 generations.

Obviously, the fixation curves lie always above the corresponding search curves. All computed fixation curves bend into a U-shape, indicating that even for common structures fixation is only possible in an interval for the mutation rate. Nevertheless, as for the search curves, there are significant differences between rare and common structures. While for the structure of the HP2 family the fixation curve diverges approximately for $\mu \approx 0.08$, for the rare HP6 structure, this happens for $\mu \approx 0.02$. This clearly shows that the critical mutation rate is not only a property of sequence length, but strongly depends on phenotype.

5. Dependence of search times on population size and evolutionary algorithm

Before exploring the dependence of the search time on the motif frequency, we want to understand to which extent the specific evolutionary algorithm and the system size influence the results. To this end, we investigate three variations of the evolutionary algorithm described above. Using still the probability function

$$p(d_i) = Z^{-1} \exp(-\beta d_i), \tag{2}$$

we consider the following modifications: (a) Scaling of the selective pressure parameter β. It may be constant, like $\beta_1 = 2/n$ as above, or time-dependent, e.g., $\beta_2 = 1/d$, where d is the average distance of the population to the target

Accepted manuscript

ACCEPTED MANUSCRIPT

structure (Stich et al., 2007). (b) Instead of using the base-pair distance d^{bp}, we use the Hamming distance d^{H}. The Hamming distance can only be measured between molecules of the same length, since it results from a positionwise comparison of the structural state. Two molecules may have a large base-pair distance and a significantly smaller Hamming distance. The first is actually a better measure of the number of changes to be performed to go from one structure to another. As an example, structures `((((....))))` and `((((..))))..` are at a base-pair distance of 8 but at a Hamming distance of 2, while structures `((........))` and `.(((....)))..` are at a Hamming distance of 6 and at a base-pair distance of 3. In this sense, the two distance measures do not value in the same way the similarity with the target and thus perform differently along the selection process (see below). Tree-edit distance has been used for comparison, needing longer evolution and simulation times and has not been utilized further.

*** Figure 2 NEAR HERE ***

In Fig. 2(a) we show for the four algorithms (besides the standard $(\beta_1, d^{\mathrm{bp}})$ also $(\beta_1, d^{\mathrm{H}})$, $(\beta_2, d^{\mathrm{bp}})$, and $(\beta_2, d^{\mathrm{H}})$), and for a specific target structure and mutation rate, how the search time g varies with the population size N. All curves are decreasing, i.e., regardless of the chosen algorithm, a given target structure is easier to find if the population is large. This is not surprising since a large pool contains with a higher probability a sequence folding relatively close to the target structure, and hence offers the population the possibility to advance faster. For both distance measures, the faster algorithm is the one with time-dependent scaling which is probably due to the fact that $\beta_2 > \beta_1$ whenever $d < n/2 = 17.5$ which is usually fulfilled after the initial stages of the evolution (and, depending on the target structure and distance measure, even already for the initial random population). We also observe that using the base-pair distance, the search process is almost one order of magnitude faster. This may be due to the fact that a sequence can have a structure with a small Hamming distance to the target, while the number of mutations necessary to yield a sequence whose minimum free energy structure coincides with the target structure may be much larger. On the contrary, if the base-pair distance is small, also the number of necessary mutations is generally small. Nevertheless, this may also depend on the considered target structure.

A different constant value of β, other than $\beta_1 = 2/n$, obviously changes the selective pressure and hence the velocity of the evolutionary process, as shown in Fig. 2(b).

The main result of this section is that the evolution algorithm is robust and that the qualitative behavior of the dependence of the search time on frequency does not depend on algorithmic details.

6. Search times as function of the motif frequency

In the sections above, we have shown how the search time depends on mutation rate, population size, evolutionary algorithm, and selective pressure parameter, and therefore we fix now the corresponding parameters and explore how

Accepted manuscript

ACCEPTED MANUSCRIPT

the frequency of a given motif (or secondary structure) influences the search time.

*** Figure 3 NEAR HERE ***

In Fig. 3, we show how g varies as a function of the frequency of appearance of each structure in the random pool (see Table 1). We observe that the more common a given secondary structure (motif) is, the faster is it found in the evolutionary process. This means that the frequency, a property purely related to the sequence–structure map, and a plain consequence of the folding landscape, clearly correlates to the velocity with which a structure is found in an evolutionary process. In other words: sequences with simple motifs are not only more frequent but also easier to find through selection (in the relevant case when they are not present in the initial population and have to be found through mutation and selection of that starting pool).

The lowest data points displayed correspond to the MAS of the SL, then come the HP, HP2 and DSL families. As anticipated in Fig. 1(a), the most frequent structures have the shortest search times. The other structures with $f > 1$ have intermediate search times, in general increasing as f decreases. With $f = 1$, we present structures belonging to classes HP6, DSL5, and HH (the sHH structure). There, the values of g spread a lot which may be due to at least two effects. First, rare structures may be composed of a large number of short loops and stems and may be very biased in their nucleotide composition, and therefore only accessible to few sequences. Second, since frequencies were determined by folding random sequences, the observed frequency $f = 1$ among 10^8 randomly sequences may over- or underestimate the actual frequency (that could only be obtained by exhaustive enumeration of sequences or estimated by inverse folding).

To give a quantitative measure of the efficiency of the evolutionary search, we have added in Fig. 3 a curve corresponding to the number of generations g_r required to find a structure of frequency f in the limit $\mu \to 1$ (or in a random search without selection), $g_r = (M/N)f^{-1}$, with $M = 10^8$ being the size of the random pool used to determine f, and N the population size. Except for the most abundant structures in the SL and HP families, all search times are below the random expectation g_r.

7. Search times as function of the number of base pairs

In the following, we study in more detail how the search time g depends on the total number of base pairs of the secondary structure. Our results are summarized in Figure 3(b), where we show g as a function of the frequency f for four different numbers of base pairs. To avoid possible effects from other structural elements, all selected structures belong to the HP family and have a hairpin loop of size four. Then, we have drawn the structures randomly across the frequency range, 43 structures in total. For example, the black curve is computed by calculating g for 11 structures with 6 base pairs in total and frequencies ranging from 3 through 6607. Again, we observe a decrease of g as the frequency of the structure increases. This decrease is unambiguous and

Accepted manuscript

ACCEPTED MANUSCRIPT

also found for other numbers of base pairs (other curves), and hence confirms the results shown above for a larger sample. As in Fig. 3(a), we represent the expectation of a random search, $g_r = 5 \times 10^5 f^{-1}$ (in these examples $N = 500$), and obtain a similar performance of the evolutionary process. Since it seems that for larger number of base pairs the values of g are systematically lower, we explore the dependence on the number of base pairs explicitly.

*** Figure 4 (new) NEAR HERE ***

In Figure 4, we show how the search and search-plus-fixation times depend on μ for three structures of *equal* abundance in a random pool ($f = 50$). A strong difference between the structures is observed. The only difference lies in the number of base pairs since the structures have not only equal abundance, but are also HP structures with a tetraloop. The black curve shows that the search time for an HP structure with 13 base pairs is found quickly with a weak dependence on μ, and fixed efficiently over a range of $0.002 \leq \mu \leq 0.06$, an HP structure with three base pairs is fixed only in a narrow interval of $0.05 \leq \mu \leq 0.15$.

*** Figure 5 NEAR HERE ***

To show that this behavior is more general, we study in Fig. 5(a) for a larger sample of structures (120 structures from the HP family) how the search time g decreases with the number of base pairs. To minimize the influence of the frequency of the structure and a possible effect of the loop size, we have considered only structures with frequency 50 and hairpin loop of size four. From this figure we can conclude that g decreases with increasing number of base pairs. A qualitatively similar behavior has been observed for a sample of structures of the DSL and HP2 families (results not shown). This result points to a non-trivial distribution in sequence space of structures with certain motifs. Actually, we know that stable structures with few base pairs show a nucleotide composition significantly far from 25% per type of nucleotide (Stich et al., 2008). On the other hand, structures with long stacks (which are usually much more stable) admit more variation in their composition. Thus, the decrease in g with number of base pairs is reflecting the spread of the corresponding neutral network in sequence space. In the next section, we explore the relation between search time and composition in more detail.

In Figure 5(b), we display g as a function of the loop size, keeping the total number of paired bases constant. We consider 120 HP structures with frequency 50 and a total number of base pairs of 7 (including the 16 structures with loop size 4 from (a)). No significant dependence on the loop size is observed.

8. Search times as function of the nucleotide composition

Another parameter that influences the sequence–structure map is the nucleotide composition of the underlying sequences. For example, rare structures have on average a bias towards a large content of G (Stich et al., 2008). Since G favors the formation of strong base pairs, short stems appear with higher probability, which in turn often build structures with less common structural

Accepted manuscript

ACCEPTED MANUSCRIPT

elements. When a structure presents restrictions in the possible composition compatible with it, it immediately limits the amount of possible different sequences. Suppose an ensemble of molecules formed by all those sequences containing a fraction f_K of each type of nucleotide, $K = \{\text{A, C, G, U}\}$. The number N_s of different sequences is

$$N_s = \frac{n!}{\prod_K \Gamma[f_K n + 1]}\,, \tag{3}$$

where $\Gamma[x]$ is the Gamma-function (that extends the factorial function to non-integer values) and n is the length of the sequences. Deviations in the composition from 25% for each nucleotide cause decreases in N_s, i.e., the sequence diversity of the ensemble diminishes. For example, an ensemble of sequences of length $n = 35$ with all four nucleotides equally present ($N_s^{(1)} = 5.59 \times 10^{18}$) is almost 5.000 times more diverse in terms of sequences than an ensemble formed equally by A, C, and G, but not containing U ($N_s^{(2)} = 1.16 \times 10^{15}$).

Our previous analyses (Stich et al., 2008) showed that the relative nucleotide content of the sequences folding into a particular structure family deviates from 25% significantly for practically all structure families. Hence, we expect that by tuning the mutation rates of the four nucleotides, the search process may be also improved or hampered. The size of the space of attainable sequences is in turn modified, as explained above.

*** Figure 6 NEAR HERE ***

In Fig. 6, we show for three target structures, a common, a rare, and one with an intermediate g value, how the nucleotide composition influences the search time, comparing always with the case of 25% of mutation probability for all types of nucleotides, A, C, G, and U (Fig. 1). First, we consider the MAS of the HP2 family, a common structure, and a distribution of mutation probabilities of $(f_A, f_C, f_G, f_U) = (0.23, 0.26, 0.30, 0.21)$, the average composition of rare structures. The resulting curve for $g(\mu)$ shows that there is no effect of the nucleotide bias on the search time for this structure. The same nucleotide bias is now applied to a rare structure, the MAS of the DSL5 family. This time, the g values are significantly lower and hence the search process is more efficient. A comment on the actual diversity of sequences attainable through this procedure is here due. Equation (3) yields the number of sequences obtained with a fixed composition of nucleotides. The probabilistic approach used in the evolutionary process permits deviations from that average which are of the order of $\sqrt{n}$ (and consequently lead to changes of possible mutation probabilities f_K of the order of $1/\sqrt{n}$), that is, they are especially important for short sequences. The actual diversity explored by the population is proportional to N_s and we can compare two populations searching around different average compositions by the ratio of the corresponding values N_s.

Finally, we consider the MAS of the HH2 family which without bias took intermediate search values. We here apply two different nucleotide compositions: one corresponds to the average composition of the 7 sequences that fold into that structure (data from (Stich et al., 2008)) and is called adjusted, f_{adj}. The other

Accepted manuscript

ACCEPTED MANUSCRIPT

one, called deplaced, corresponds to a composition $f_{\rm dep} = 0.5 - f_{\rm adj}$ (restricting ourselves to values $f_{\rm adj} \leq 0.5$). This is just an example. Obviously, there is no unique way of deplacing a composition from the optimal one. The adjusted composition has mutation rates of $(f_A, f_C, f_G, f_U) = (0.22,0.26,0.32,0.20)$, and the deplaced one rates of $(f_A, f_C, f_G, f_U) = (0.28,0.24,0.18,0.30)$. The curves clearly demonstrate that an adjusted composition decreases the search time, while a deplaced one increases the search time with respect to the neutral composition, thus clearly revealing the position occupied by the neutral network of that structure in sequence space. In particular, for large mutation rates, the difference in g values between a favorable and an unfavorable composition may be of one order of magnitude in g, and similar g values are obtained for mutation rates differing in a factor of two. Although the displayed curves just represent a few examples, we expect the effect to be generic, although, in any case, the optimal composition is structure-dependent. Regarding the number of attainable sequences N_s, the even composition of 25% yields in this case 3.63 times more different sequences than the adjusted composition and 1.78 times more than the deplaced one. The fact that g is lowest for the adjusted case in spite of the low number of attainable sequences N_s, clearly demonstrates the importance of an appropriate nucleotide bias for the search process.

9. Summary and Discussion

Summary. The results presented in this article relate properties of the sequence–structure map, which are completely determined by the underlying RNA folding process, with properties of an evolving population, in particular the search and search-plus-fixation time. We have shown in this article that the abundance of a structure in a random pool is an important parameter that determines the search process. Nevertheless, it is clearly not the only one: the motif and the composition of the sequence (nucleotide bias), together with the structure abundance determine the search process.

For a set of 18 structures representing the relevant structure families for molecules of length 35 nt, the evolutionary accessibility has been shown in Fig. 1(a): structures which are common in a random pool are found faster than rare structures for the range of relevant mutation rates. This result probably could have been expected since frequent structures have a larger set of sequences folding into the same structure and hence the possibility that a population finds any of the sequences is higher.

Figure 1(a) confirms that rare structures are not only found much more slowly, but also that the search is only effective in a small range of the mutation rate. This finding has important consequences for the evolvability of RNA structures: It means that unless the mutation rate is sufficiently low, finding specific rare structures is difficult. For inappropriately high mutation rates, the population will be typically composed of simpler and more frequent structures, and rare structures will go unnoticed. This effect becomes even more pronounced if we consider the search-plus-fixation time (Fig. 1(b)), which is the quantity that describes evolutionary success within the framework of this model (Stich et al.,

Accepted manuscript

ACCEPTED MANUSCRIPT

2007). The fact that different structures, even of the same length, do not necessarily have the same search times, search-plus-fixation times, or identical values of the asymptotic values of the population (like d or the fraction of correctly folded molecules), has been already shown before (Stich et al., 2007). There, however, only two different structures were considered, which also behaved relatively similar, while here we have worked with 18 structures in detail and have been able to show that the ranges of mutation rates permitting search and fixation may differ strongly. This result adds a new variable to the parameters that influence the critical mutation rate for a phenotype to be present in a population. Up to now, neutrality (that is the existence of many different genotypes yielding the same phenotype) has been discussed as an important parameter in determining the mutation rate that a population can bear without losing its master phenotype, or its viability. We here show clearly that rare phenotypes are at the same time more difficult to find and much more difficult to maintain in the population.

In Figure 3, we presented for the 18 representative structures how for a given mutation rate the search time depends on the frequency. These structures cover the whole accessible frequency range (more than 5 orders of magnitude) and have search times varying on more than 3 orders of magnitude. We show therefore convincingly that frequent structures are faster to find and confirm the findings already presented in Fig. 1.

To clarify to which extent the results depend on elements of the secondary structure such as loop size and number of base pairs, we performed additional simulations for 375 other structures. Figure 3(b) clearly showed that the findings of Fig. 3(a) are generic and by no means special to the 18 structures first chosen, and that for practically the whole frequency range, structures with more paired bases are easier to find. To show this more clearly, in Fig. 4 we have displayed for three structures of *equal* abundance that the search and search-plus-fixation process is much more efficient for structures with many base pairs, thus showing that evolvability of structures goes beyond mere abundance. The latter finding has been further investigated and in Fig. 5 we have studied how the search time depends on the total number of base pairs and on the loop size. While the number of base pairs influences strongly the search time, the loop size does not.

Finally, we have investigated the influence of the nucleotide composition on the search time, considering mild biases from the equiprobable distribution of 25% for all types of bases. As shown in Fig. 6, a bias towards an unfavorable (favorable) nucleotide composition for a given rare structure leads to an increase (decrease) of the search time. However, for frequent structures, a mild bias does not show any significant effect on the search times.

Neutral networks. Many of the results of this article can be interpreted in terms of the structure and size of neutral networks. If all sequences folding into the same secondary structure are related to each other by single mutational events, the network is said to be connected, while it is also possible that the network is disconnected and comprises different components which are at a mutational distance larger than one. Rare structures have smaller neutral networks which furthermore may be disconnected and located in some restricted areas of

Accepted manuscript

ACCEPTED MANUSCRIPT

sequence space. The abundance of a given structure could have conditioned its appearance in natural RNA, functional motifs, as studies with molecules up to 18 nt in length have shown (Cowperthwaite et al., 2008), further demonstrating that rare motifs are difficult to access through evolution on neutral networks of common motifs. That is, they are not homogeneously distributed in sequence space. As a consequence, a bias in nucleotide composition may slow down or accelerate the search process significantly, as has been reported above: accessibility depends on the composition of the pool and on mutations happening with different probability to each of the nucleotides. On the other hand, one can assume that frequent structures have neutral networks that are connected and that are distributed relatively homogeneously across the sequence space, such that accessing them is independent of composition and of uneven mutation rates. The results presented above agree with other studies where the impact of nucleotide composition on the structural repertoire has been studied in detail (Knight et al., 2005; Kim et al., 2007).

Sequence length and evolvability. Since our work is based on the computational determination of secondary structures and *in silico* evolving RNA populations, we are limited in the sequence length that we can consider. What effects can be expected in the case of longer sequences? Obviously, the numbers of sequences and structures increase, although the ratio of the size of the structure space to the sequence space decreases. Furthermore, for longer sequences, the fraction of sequences folding into common structures compared to rare structures becomes larger (Grüner et al., 1996a). Therefore, we expect that the search time for a rare structure increases compared to the search time for a common structure. A very general feature of replicating populations is that there exists a critical mutation rate above which the population cannot maintain function. This critical mutation rate is closely related to the sequence length: if selection acts upon the genotype, this threshold is inversely proportional to the length of the sequence. If selection acts upon the phenotype, higher mutation rates are allowed. In our simulations, this critical mutation rate is given by the fixation threshold. Our results show that this fixation threshold varies strongly with the structure and we clarified that besides abundance in particular the number of base pairs and composition of the random pool affect this threshold. It turns out that common structures are not only easier to find, but also withstand higher mutation rates and are therefore more evolvable. Although our simulations were not aimed to explain any specific scenario, the molecule length used here ($n = 35$) leads to critical error rates which are compatible with a recent experimental study for nonenzymatic nucleic acid replication (Rajamani et al., 2010).

Other folding landscapes. In this article, we used as phenotype the RNA secondary structure as minimum free energy structure, as predicted by the `fold()` routine of the Vienna RNA package (see Methods). Our aim is not to perform an accurate prediction of an RNA structure from a given sequence –a goal which should also take into account tertiary interactions or, even in the realm of secondary structure prediction, suboptimal structures, partition functions, or kinetic folding–, but to obtain statistically significant properties of the evolu-

Accepted manuscript

ACCEPTED MANUSCRIPT

tionary search for a particular phenotype. To our advantage, it has been shown that structure statistics is relatively insensitive to the precise folding algorithm used (Tacker et al., 1996). There, in particular the autocorrelation function of structures, the existence and component structure of the neutral network, and the shape space covering property were shown to be basically independent of the folding algorithm and its parameters. We hence expect that the qualitative behavior of the average search time, computed from 200 independent realizations from populations of size 500-10000, is qualitatively independent of the details of the folding algorithm. The investigation of a model of replicating RNA molecules where selection took into account the suboptimal structures, has been carried out by Ancel and Fontana (2000). The resulting *plastic* repertoire of structures of a sequence contributes to the overall fitness of the sequence in proportion to its weight in the Boltzmann distribution. As a result of selection, this repertoire becomes smaller, and the reduction of this plasticity (repertoire of structures) goes hand in hand with a loss of variability and hence loss of evolvability. This is explained by the correlation of the set of attainable (MFE plus suboptimal) structures of a given sequence with the MFE structures of the genetic neighborhood of that sequence. This statistical property of the RNA genotype–phenotype map is responsible for an evolving population to approach areas of sequence space with high neutrality. Our work agrees qualitatively with these findings, and furthermore shows that the loss of evolvability is differential, i.e., depends on the structure used as target of evolution and in particular on the composition of the nucleotide pool and the structural motifs.

Fitness landscapes. Using RNA secondary structures as target of evolution, the fitness landscape is rough and its properties can be characterized in terms of autocorrelation functions (Fontana et al., 1991, 1993; Stadler, 1999). For the model studied by Ancel and Fontana (2000), results were reported to be robust with respect to changes in the functional form (hyperbolic, exponential, linear) of the fitness function. Here, we apply an exponential fitness function, which is believed to be more realistic than fitness functions varying linearly with the structural distance (Bonhoeffer et al., 1993). This fitness function is rather general since it includes as limit cases the situation where all structures have identical fitness ($\beta = 0$) and that where only the target structure has nonzero fitness ($\beta \to \infty$). Considering a wide range of different secondary structures as target structures, and varying selective pressures and selection algorithms, we have studied the effect of different fitness landscapes on the evolutionary dynamics.

Outlook. The results for our evolutionary model show that the search for a common structure, and hence of its associated chemical function, is much more efficient (faster) and reliable (in a wider range of mutation rates) than the search for a rare structure. Eventually, we would like to establish a correspondence between a set of sequences and the chemical functions they can perform. In the same way that the sequence–structure map offers a huge number of sequence solutions for a fixed structure, it might be that the structure–function map is redundant to the point of permitting common structures (maybe with additional requirements, as bearing particular sequences or having low folding

Accepted manuscript

ACCEPTED MANUSCRIPT

energy) to perform "rare" functions. If this is so, the appearance of simple RNA motifs in natural, functional molecules (such as reported in (Gan et al., 2003; Cowperthwaite et al., 2008)), would not be just a contingent fact due to their high frequency in random pools, but a consequence of the complete sequence–structure–function relationship. This is just a hypothesis to keep in mind, since the relation between molecular structure and chemical function (which is the actual phenotype of a molecule) is an open field of study. Two counteracting forces are present: the greater abundance and evolvability of common structures versus the possibly better performance of some specific rare structure.

The relationship among nucleotide composition, structural motifs, and size and connectedness of neutral networks appears as a subject deserving further research. Advances in these areas might have potential applications to *in vitro* evolution of short sequences (ribozymes, RNA and DNA aptamers, or other polymeric ligands) and, in the context above, to start disentangle the relationship between the genomic restrictions of a given phenotype and its evolutionary attainability.

10. Methods

Simulations have been carried out at the Itanium II cluster of INTA (Instituto Nacional de Técnica Aeroespacial, Spain). For random number generation, we relied on the Mersenne Twister algorithm as provided by GNU Scientific Library (GSL), Version 1.7 (see http://www.gnu.org/software/gsl). For secondary structure folding (minimum free energy) and calculation of base-pair and Hamming distances, we use the Vienna RNA package (Hofacker et al., 1994), version 1.5, with the current standard parameter set and allowing for A-U, G-C, and G-U base pairs, and the formation of isolated base pairs. The list of secondary structures used as target structures can be obtained from the authors (besides those given in Table 1).

The search (fixation) times are determined as average values over $R = 200$ ($R = 20$) independent simulations (only in Fig. 4 for $R = 100$). For reasons of clarity, standard deviations are given as error bars in Figs. 2, 3 and 6 only. In Fig. 5 we show average search times obtained for various structures. For each structure, the average was obtained individually as above ($R = 200$), and then average over structures was performed. The error bars there indicate the standard deviation over the ensemble of structures.

Accepted manuscript

Acknowledgements

Author contributions. All authors conceived and designed the study. MS performed the numeric calculations. All authors analysed the data and wrote the paper.

Funding. The authors acknowledge support from Spanish MICIIN through project FIS2008-05273 and from Comunidad Autónoma de Madrid, project MODELICO (S2009/ESP-1691).

ACCEPTED MANUSCRIPT

References

Ancel, L. W., Fontana, W., 2000. Plasticity, evolvability, and modularity in RNA. J. Exp. Zool. 288, 242–283.

Bartel, D. P., Szostak, J. W., 1993. Isolation of new ribozymes from a large pool of random sequences. Science 261, 1411–1418.

Bonhoeffer, S., McCaskill, J. S., Stadler, P. F., Schuster, P., 1993. RNA multi-structure landscapes. Eur. Biophys. J. 22, 13–24.

Briones, C., Stich, M., Manrubia, S. C., 2009. The dawn of the RNA world: Toward functional c omplexity through ligation of random RNA oligomers. RNA 15, 743–749.

Carothers, J. M., Oestreich, S. C., Davis, J. H., Szostak, J. W., 2004. Informational complexity and functional activity of RNA structures. J. Am. Chem. Soc. 126, 5130–5137.

Clote, P., Kranakis, E., Krizanc, D., Salvy, B., 2009. Asymptotics of canonical and saturated RNA secondary structures. J. Bioinform. Comput. Biol. 7, 869–893.

Cowperthwaite, M. C., Economo, E. P., Harcombe, W. R., Miller, E. L., Ancel Meyers, L., 2008. The ascent of the abundant: How mutational networks constrain evolution. PLoS Comput. Biol. 4, e1000110.

Dirks, R. M., Lin, M., Winfree, E., Pierce, N. A., 2004. Paradigms for computational nucleic acid design. Nucleic Acids Res. 32, 1392–1403.

Fontana, W., Griesmacher, T., Schnabl, W., Stadler, P. F., Schuster, P., 1991. Statistics of landscapes based on free energies, replication and degradation rate constants of RNA secondary structures. Monatsh. Chem. 122, 795–819.

Fontana, W., Konings, D. A. M., Stadler, P. F., Schuster, P., 1993. Statistics of RNA secondary structures. Biopolymers 33, 1389–1404.

Fontana, W., Schuster, P., 1987. A computer-model of evolutionary optimization. Biophys. Chem. 26, 123–147.

Gan, H. H., Pasquali, S., Schlick, T., 2003. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. Nucleic Acids Res. 31, 2926–2943.

Gevertz, J., Gan, H. H., Schlick, T., 2005. In vitro RNA random pools are not structurally diverse: a computational analysis. RNA 11, 853–863.

Grüner, W., Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I. L., Stadler, P. F., Schuster, P., 1996a. Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks. Monatsh. Chem. 127, 355–374.

Accepted manuscript

ACCEPTED MANUSCRIPT

Grüner, W., Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I. L., Stadler, P. F., Schuster, P., 1996b. Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structures of neutral networks and shape space covering. Monatsh. Chem. 127, 375–389.

Held, D. M., Greathouse, S. T., Agrawal, A., Burke, D. H., 2003. Evolutionary landscapes for the acquisition of new ligand recognition by RNA aptamers. J. Mol. Evol. 57, 299–308.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., Schuster, P., 1994. Fast folding and comparison of RNA secondary structures. Monatsh. Chem. 125, 167–188.

Hofacker, I. L., Schuster, P., Stadler, P. F., 1998. Combinatorics of RNA secondary structures. Discr. Appl. Math. 88, 207–237.

Huynen, M. A., Konings, D. A. M., Hogeweg, P., 1993. Multiple coding and the evolutionary properties of RNA secondary structure. J. Theor. Biol. 165, 251–267.

Huynen, M. A., Stadler, P. F., Fontana, W., 1996. Smoothness within ruggedness: the role of neutrality in adaptation. Proc. Natl. Acad. Sci. USA 93, 397–401.

Joyce, G. F., 2004. Directed evolution of nucleic acid enzymes. Annu. Rev. Biochem. 73, 791–836.

Kennedy, R., Lladser, M. E., 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.

Kim, N., Gan, H. H., Schlick, T., 2007. A computational proposal for designing structured RNA pools for in vitro selection of RNAs. RNA 13, 478–492.

Knight, R., De Sterck, H., Markel, R., Smit, S., Oshmyansky, A., Yarus, M., 2005. Abundance of correctly folded RNA motifs in sequence space, calculated on computational grids. Nucleic Acids Res. 33, 5924–5935.

Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W., Haussler, D., 2006. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput. Biol. 2, e33.

Petrillo, M., Silvestro, G., Di Nocera, P. P., Boccia, A., Paolella, G., 2006. Stem-loop structures in prokaryotic genomes. BMC Genomics 7, 170.

Pitt, J. N., Ferré-D'Amaré, A. R., 2010. Rapid construction of empirical RNA fitness landscapes. Science 330, 376–379.

Accepted manuscript

ACCEPTED MANUSCRIPT

Puerta-Fernández, E., Romero-López, C., Barroso-delJesús, A., Berzal-Herranz, A., 2003. Ribozymes: recent advances in the development of RNA tools. FEMS Microbiol. Rev. 27, 75–97.

Rajamani, S., Ichida, J. K., Antal, T., Treco, D. A., Leu, K., Nowak, M. A., Szostak, J. W., Chen, I. A., 2010. Effect of stalling after mismatches on the error catastrophy in nonenzymatic nucleic acid replication. J. Am. Chem. Soc. 132, 5880–5885.

Sabeti, P. C., Unrau, P. J., Bartel, D. P., 1997. Accessing rare activities from random RNA sequences: the importance of the length of molecules in the starting pool. Chem. Biol. 4, 767–774.

Schuster, P., 2003. Molecular insights into evolution of phenotypes. In: Crutchfield, J. P., Schuster, P. (Eds.), Evolutionary Dynamics. Oxford Univ. Press, pp. 163–215.

Schuster, P., 2006. Prediction of RNA secondary structures: from theory to models and real molecules. Rep. Prog. Phys. 69, 1419–1477.

Schuster, P., Fontana, W., Stadler, P. F., Hofacker, I. L., 1994. From sequences to shapes and back: A case study in RNA secondary structures. Proc. Roy. Soc. London B 255, 279–284.

Stadler, P. F., 1999. Fitness landscapes arising from the sequence-structure maps of biopolymers. J. Mol. Struc. (Theochem) 463, 7–19.

Stich, M., Briones, C., Manrubia, S. C., 2007. Collective properties of evolving molecular quasispecies. BMC Evol. Biol. 7, 110.

Stich, M., Briones, C., Manrubia, S. C., 2008. On the structural repertoire of pools of short, random RNA sequences. J. Theor. Biol. 252, 750–763.

Stich, M., Lázaro, E., Manrubia, S. C., 2010a. Phenotypic effect of mutations in evolving populations of RNA molecules. BMC Evol. Biol. 10, 46.

Stich, M., Lázaro, E., Manrubia, S. C., 2010b. Variable mutation rates as an adaptive strategy in replicator populations. PLoS ONE 5, e11186.

Tacker, M., Stadler, P. F., Bornberg-Bauer, E. G., Hofacker, I. L., Schuster, P., 1996. Algorithm independent properties of RNA secondary structure predictions. Eur. Biophys. J. 25, 115–130.

Waldispühl, J., Devadas, S., Berger, B., Clote, P., 2008. Efficient algorithms for probing the RNA mutation landscape. PLoS Comput. Biol. 4, e1000124.

Waterman, M. S., 1978. Secondary structure of single-stranded nucleic acids. In: Studies in Foundation and Combinatorics. Vol. 1 of Advances in Mathematics Supplementary Studies. Academic Press, pp. 167–212.

Accepted manuscript

ACCEPTED MANUSCRIPT

Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess Jr., J. W., Swanstrom, R., Burch, C. L., Weeks, K. M., 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 460, 711–719.

Accepted manuscript

ACCEPTED MANUSCRIPT

Name	H	I+B	M	Structure in bracket notation	f
SL	1	-	-	((((....))))..........................	62893
HP	1	1	-	(((.(((....))).)))...................	9977
HP2	1	2	-	(((.((.(((....))).)).)))............	1275
HP3	1	3	-	(((.((.((.((....)).)).)).)))........	144
HP4	1	4	-	((.((.((.((..((....))..)).)).)).)).	16
HP5	1	5	-	((.(((.(.((.(.(....).).)).).))).)).	3
HP6	1	6	-	.(.(((.(.(.((.(.(....).)))).)))).).	1
DSL	2	-	-	((((....))))..........((((....)))).	1505
DSL2	2	1	-	((((....))))......(((.((....)).))).	190
DSL3	2	2	-	(((.((....)).)))..((.(((....))).)).	25
DSL4	2	3	-	.((((.((((....)).)).).))).((...)).	3
DSL5	2	4	-	.((.((....))))((..(.((....).).).))	1
HH	2	-	1	((((..(((....))).(((....)))..))))..	37
HH2	2	1	1	(((.((.((....)).(((....)))..)).))).	7
TSL	3	-	-	(((....))).(((....))).(((...)))...	13
TSL2	3	1	-	.(....).(((....))).(((.(....).)))..	2
sHP	1	1	-	..(((((((....(((...)))......)))))))	24
sHH	2	-	1	(((..((((....)))))..((((...)))).)))	1

Table 1: Table of the structures used in Figs. 1, 2, 3(a), and 6. The first 16 structures are most abundant structures of their respective families. The structures sHP and sHH have already been used in Ref. (Stich et al., 2007). The letter H denotes the number of hairpin loops, I stands for interior loops, B for bulges, and M for multiloops. The frequency f has been taken from the folding of 10^8 sequences (Stich et al., 2008).

Accepted manuscript

ACCEPTED MANUSCRIPT

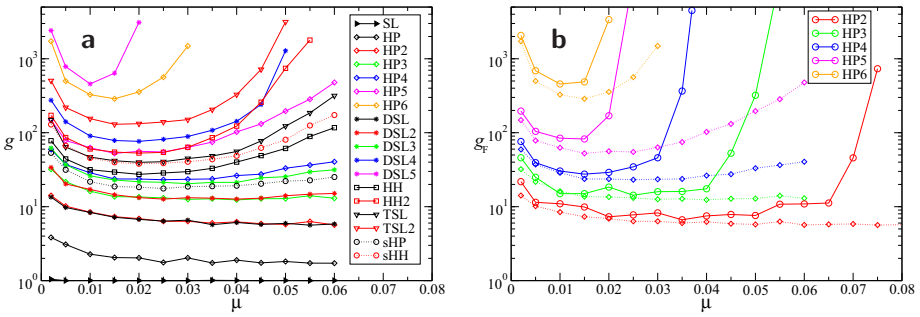


Figure 1: (a) The search time g as function of the mutation rate μ for 18 structures (16 most abundant structures (MAS) of a structure family plus the hairpin sHP and hammerhead sHH motifs considered in Stich et al. (2007)). In general, the search curve shows a U-shape, pronounced for rare structures and flattened for common ones. Also, frequent structures show significantly shorter search times. Search times have been determined from $R = 200$ independent realizations of populations with size $N = 10000$. (b) Comparison of the search times g (from (a)) and search-plus-fixation (short: fixation) times g_{F} as function of the mutation rate μ for the five structures of the structure families HP2, HP3, HP4, HP5, and HP6. The fixation curves show the typical U-shape above the search curve. For common structures, g_{F} diverges at a value of μ larger than that for rare structures. Fixation times have been determined from $R = 20$ independent realizations of populations with size $N = 10000$.

Accepted manuscript

ACCEPTED MANUSCRIPT

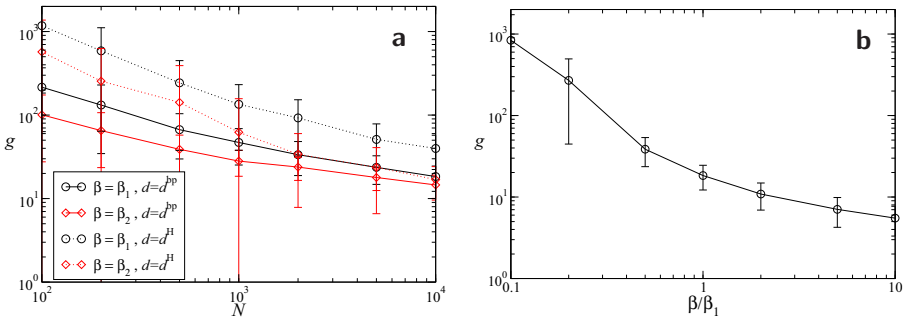


Figure 2: (a) The search time g as function of the population size N for four different evolution algorithms (see text). (b) The search time g for different strengths of the selective pressure. The target structure is the sHP structure, the mutation rate is $\mu = 0.02$, population size $N = 10000$, and the number of independent realizations is $R = 200$.

Accepted manuscript

ACCEPTED MANUSCRIPT

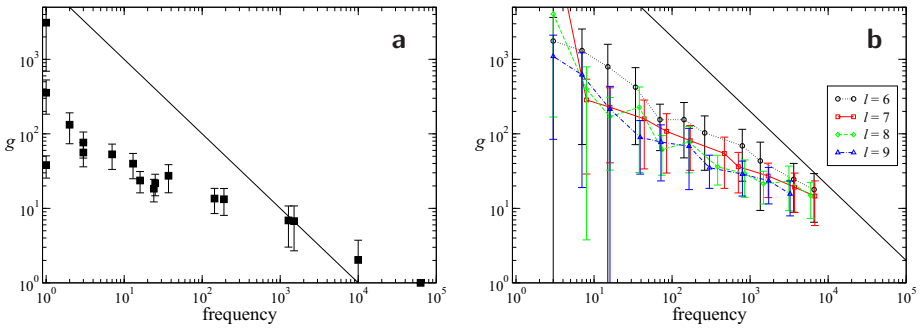


Figure 3: (a) Search time g as function of the frequency of the secondary structure for the 18 structures considered in Fig. 1. The search time clearly decreases with the frequency. The parameters are $\mu = 0.02$, $N = 10000$, and $R = 200$. (b) Search time g as function of the frequency for structures with different total number of base pairs. Here, all structures belong to the HP structure class and have a hairpin loop of size 4. The total number of base pairs l varies from 6 through 9 (see legend). Again, the search time clearly decreases with the frequency and furthermore structures with large l seem to be found faster. We have used 11 structures with $l = 6$, 11 with $l = 7$, 11 with $l = 8$, and 10 with $l = 9$, chosen randomly but with the aim of covering the whole frequency range. The parameters are $\mu = 0.02$, $N = 500$, and $R = 200$. In (a,b), the straight line stands for the average number g_r of attempts required to find a structure of frequency f within a random population of size N (see main text).

Accepted manuscript

ACCEPTED MANUSCRIPT

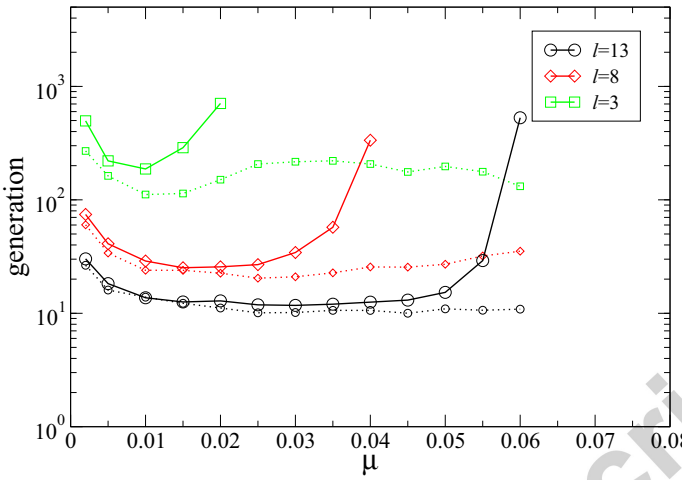


Figure 4: Search time g (dotted) and search-plus-fixation time g_{F} (solid) as function of μ for three HP structures with frequency $f = 50$, consisting of a hairpin loop of size 4 and a total number of base pairs $l = 13$ (black), $l = 8$ (red), and $l = 3$ (green). The parameters are $N = 10000$ and $R = 100$.

Accepted manuscript

ACCEPTED MANUSCRIPT

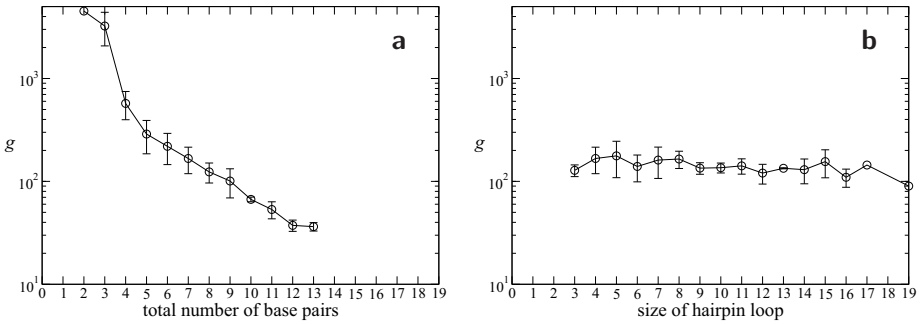


Figure 5: (a) The search time g as function of the total number of base pairs l for structures from the HP structure family. The search time clearly decreases with the total number of base pairs. We have chosen structures with frequency $f = 50$ and hairpin loop of size 4. We have used all 120 structures that fulfill the criteria: 1 for $l = 2$, 5 for $l = 3$, 10 for $l = 4$, 16 for $l = 5$, 19 for $l = 6$, 16 for $l = 7$, 14 for $l = 8$, 13 for $l = 9$, 7 for $l = 10$, 5 for $l = 11$, 9 for $l = 12$, and 5 for $l = 13$. (b) The search time g as function of the hairpin loop size p for all 120 structures from the HP familiy with frequency $f = 50$ and total number of base pairs $l = 7$. 17 for $p = 3$, 16 for $p = 4$ (already considered in (a)), 18 for $p = 5$, 6 for $p = 6$, 15 for $p = 7$, 10 for $p = 8$, 9 for $p = 9$, 4 for $p = 10$, 5 for $p = 11$, 6 for $p = 12$, 2 for $p = 13$, 4 for $p = 14$, 3 for $p = 15$, 3 for $p = 16$, 1 for $p = 17$, and 1 for $p = 19$. No significant dependence is observed. (a,b) The parameters are $\mu = 0.02$, $N = 500$, and $R = 200$.

Accepted manuscript

ACCEPTED MANUSCRIPT

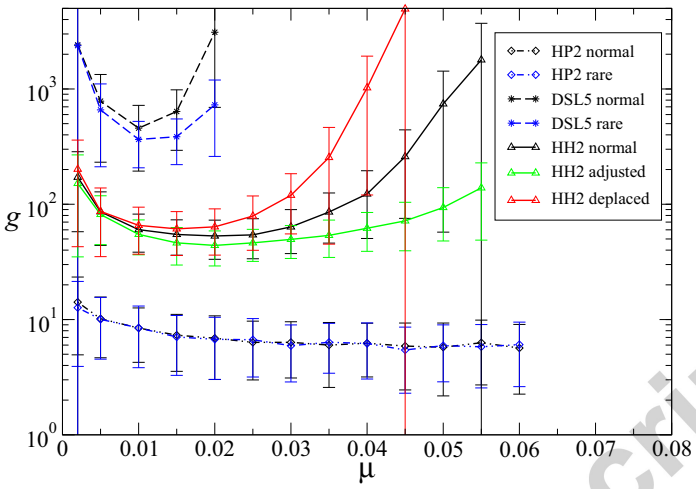


Figure 6: Search time g as function of μ for the structures representing families HP2, HH2, and DSL5 as the nucleotide composition is varied. The population size is $N = 10000$ and $R = 200$ independent realizations were performed.

Accepted manuscript

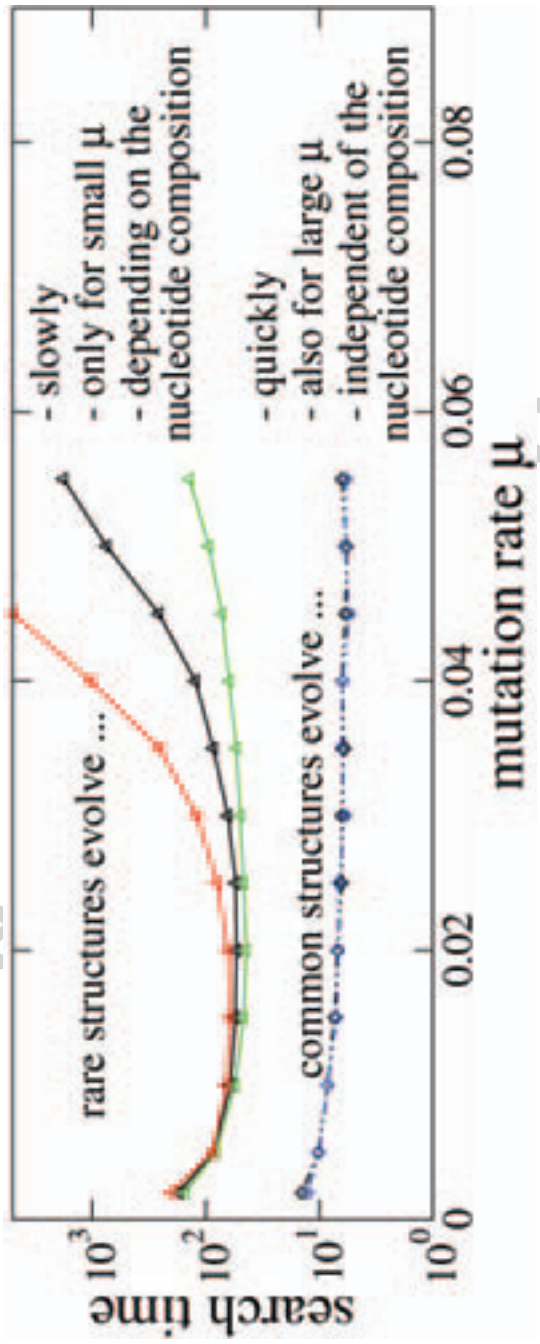
ACCEPTED MANUSCRIPT

Research highlights:

* evolvability of RNA structures depends on motif frequency
* evolvability of RNA structures depends on the type of motif
* rare structures evolve only in a narrow interval of the mutation rate
* evolvability of rare RNA structures depends on the nucleotide composition

Accepted manuscript

ACCEPTED MANUSCRIPT



Accepted Manuscript