Modular evolution and increase of functional complexity in replicating RNA molecules

SUSANNA C. MANRUBIA and CARLOS BRIONES
Centro de Astrobiología, CSIC-INTA, Torrejón de Ardoz, Madrid, Spain

ABSTRACT

At early stages of biochemical evolution, the complexity of replicating molecules was limited by unavoidably high mutation rates. In an RNA world, prior to the appearance of cellular life, an increase in molecular length, and thus in functional complexity, could have been mediated by modular evolution. We describe here a scenario in which short, replicating RNA sequences are selected to perform a simple function. Molecular function is represented through the secondary structure corresponding to each sequence, and a given target secondary structure yields the optimal function in the environment where the population evolves. The combination of independently evolved populations may have facilitated the emergence of larger molecules able to perform more complex functions (including RNA replication) that could arise as a combination of simpler ones. We quantitatively show that modular evolution has relevant advantages with respect to the direct evolution of large functional molecules, among them the allowance of higher mutation rates, the shortening of evolutionary times, and the very possibility of finding complex structures that could not be otherwise directly selected.

Keywords: RNA folding; molecular evolution; genotype–phenotype relationship; structural motifs; RNA world

INTRODUCTION

The size of the first informative molecules was strongly constrained by the accuracy in replication. In an environment where proofreading mechanisms were initially absent, replicating biomolecules had to be necessarily short. This represented a strong limitation in the amount of genetic information that could be stored and reliably transmitted to subsequent generations, as well as to the functional capabilities of the evolving molecules. That process likely led to the appearance of molecular quasispecies (Eigen 1971), large and heterogeneous populations of replicating molecules that initiated Darwinian evolution.

One of the most popular scenarios for molecular evolution prior to the appearance of cellular life is that of the RNA world (Gilbert 1986; Joyce 2002), where small populations of replicating RNA molecules would simultaneously encode information and perform catalytic activity. Mutation (inherent to the replication process) and recombination should have promoted the appearance of variants. Selection, defined through the characteristics of the environment where evolution proceeded, would have favored the replication of certain molecular types. Different microenvironments (characterized by their physicochemical conditions, including ionic strength, pH, metal concentration, or temperature) would then induce different selection pressures, and eventually a spectrum of independent populations of functional replicators might have been simultaneously available. In a favorable situation, it is possible that each molecular quasispecies selected in that way specialized in performing a single, simple function, as a step prior to the emergence of genetic or metabolic reaction networks. This scenario has received steadily increasing experimental support in the last two decades.

Although there is no known natural ribozyme that catalyzes the template-directed polymerization of nucleotides, in vitro evolution experiments have shown that RNA-dependent RNA polymerization can be performed without the help of proteins (Johnston et al. 2001; Joyce 2004; Orgel 2004). However, the details of how such a process could have taken place in the RNA world are as yet unknown (Joyce 2002; Joyce and Orgel 2006). Advances in experimental research indicate that the appearance of complex functions in RNA molecules could have been linked to the independent selection of molecular motifs or domains rather than to the de novo selection of complex molecules (Knight and Yarus 2003; Joyce 2004; Wang and Unrau 2005). Indeed, it has been experimentally shown that catalytic RNAs can be enriched with new functional abilities using as a starting point certain domains of pre-existing

Reprint requests to: Susanna C. Manrubia, Centro de Astrobiología, CSIC-INTA, Ctra. de Ajalvir km. 4, 28850, Torrejón de Ardoz, Madrid, Spain; e-mail: cuevasms@inta.es; fax: 34-91-5206424.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.203006.

natural (Jaeger et al. 1999) or in vitro evolved (Johnston et al. 2001) ribozymes, appended to random-sequence segments. Also, functional motifs have been successfully combined to generate allosteric ribozymes (Tang and Breaker 1997; Komatsu et al. 2002), effector-activated ribozymes (Robertson and Ellington 2000), and complex molecules endowed with two activities, such as RNA cleavage and ligation (Landweber and Pokrovskaya 1999; Kumar and Joyce 2003). More recently, a two-step in vitro evolution method was developed that allowed the sequential selection for specific ligand binding and cleavage, so that the evolved catalytic RNAs carried an aptamer domain and a ribozyme one (Romero-López et al. 2005). A different experimental approach was carried out to evolve minimal functional RNAs, leading to the characterization of certain activities performed by individual RNA modules (Lozupone et al. 2003; Wang and Unrau 2005). As an example, it was demonstrated that fragmented ribozymes are able to catalyze RNA ligation reactions (Vlassov et al. 2004). Complementary research on extant RNA has addressed the existence of functional domains in different RNA molecules, and the ensuing relation between sequence, structure, and function (Westhof and Massire 2004; Lilley 2005; Noller 2005).

Relevant in the appearance of a certain function is the ability of an evolving population of RNA sequences to find a solution (a given structure, and subsequently a given function) in the space of sequences. The map between sequence and structure then becomes a main issue, since there are many different RNA sequences folding into the same secondary structure. The set of all those sequences forms the so-called neutral network corresponding to the structure. Theoretical research on neutral networks suggests that there are sequences arbitrarily close to any pair of structures. Experimental confirmation of this thesis arose recently, with the demonstration that certain RNA sequences can fold, with minor changes, into two unrelated secondary structures endowed with two different catalytic capabilities (Schultes and Bartel 2000; Held et al. 2003). These experimental observations thus imply that it is possible to evolve a single RNA sequence such that it performs different functions corresponding to different structural conformations. The change of conformation could be triggered through minor environmental changes, which suggests interesting evolutionary implications for potentially bifunctional ribozymes.

All these experiments notwithstanding, little is known about the gap that expands between the longest nucleic acid molecules that can be obtained nonenzymatically (Luther et al. 1998; Huang and Ferris 2003) and the shortest functional RNA molecules that can perform the complex activities required for the establishment of an RNA world (Orgel 2004; Joyce and Orgel 2006). Furthermore, a detailed analysis of the structure–function relationship underlying modular evolution is still lacking, both in the field of the early evolution of genetic machinery and in the design of RNA evolution experiments. Both computer (in silico) simulations and experimental (in vitro) work are needed in order to understand how the increase of molecular size and complexity could have driven the emergence of biological functions from a chemical world.

RNA secondary structure offers an appropriate representation of the genotype–phenotype map (Schuster et al. 1994; Fontana and Schuster 1998; Ancel and Fontana 2000; Fontana 2002). It is realistic yet simple enough to address evolutionary questions whenever a separation between genotype (object of mutations) and phenotype (object of selection) is required. Basic evolutionary concepts like the degeneration of structure spaces (Huynen 1996); the neutral drift in a population; the existence of a phenotypic error threshold (Huynen et al. 1996; Kun et al. 2005; Takeuchi et al. 2005); the evolution toward maximally connected regions in the neutral space of each secondary structure (van Nimwegen et al. 1999; Wilke 2001); or the connection between neutral networks, implying closeness of almost any pair of secondary structures (Schuster 1993), arise in this framework in a natural way.

Here we analyze, using computational simulations and theoretical approaches, the combination of different RNA modules in the search for a complex functional molecule. Our aim is to quantify the likelihood that two short, independently evolved sequences could combine and give rise to a molecule endowed with their two original functions. Two different functionalities (phenotypes) of RNA molecules are represented as two different secondary structures, each one standing in its turn for the optimal conformation in the environment where the corresponding population evolves. We start with two small pools of short, random sequences that independently evolve through a large number of replication cycles under the action of point mutations and selection. In a wide range of mutation rates, sequences folding into the target structure are eventually found and fixed in either population. Then we allow the two populations to mix such that, under certain conditions, molecules are ligated. Secondary structures corresponding to those sequences that are twice as long occasionally retain the structures of the two original modules. This allows us to propose a scenario in which the combination of previously evolved RNA modules can produce longer and functionally more complex molecules. We quantitatively show that this modular evolution model is much more likely than the alternative one: a direct evolution in which complex molecules would emerge with the two modules and functions at once. Our results have implications for the understanding of the evolutionary pathways that allowed an increase of functional complexity in the first macromolecules combining genotype and phenotype in the context of an RNA world.

RESULTS

The results that we present in this section have been obtained from computer simulations in which the evolution

and selection of a population of RNA sequences are numerically implemented. Nevertheless, we use realistic parameters derived from in vitro experiments in order to obtain not only qualitative, but also quantitative results that may be relevant for future experimental work.

Model and implementation

Evolution of a single population

The simulation of the evolution of each independent population starts with a pool of $N = 602$ (1 zeptomole) random sequences of length $n = 35$, a reasonable population size and molecular length in the context of an RNA world (Knight and Yarus 2003). At each time step, the minimum free-energy secondary structure corresponding to each sequence is calculated (see Materials and Methods). We consider two independent populations (A and B) of RNA sequences that evolve according to the rules specified. We have chosen two target structures analogous to two of the shortest catalytic RNAs found in nature. Previous experimental work has shown that both ribozymes are involved in the replication processes of the RNA genomes containing them, and their catalytic domains have been in vitro engineered to support *trans*-cleavage reactions (Doudna and Cech 2002; Puerta-Fernández et al. 2003; Lilley 2005).

The target structure assigned to population A is a hairpin-like structure, analogous to that of ribozymes that induce specific cleavage of the satellite RNAs in which they are found (Hampel and Tritz 1989). Furthermore, they are able to perform RNA ligase activities (Buzayan et al. 1986; Prodi et al. 1986). The second set of sequences (population B) evolves toward a hammerhead motif, similar to that found in self-cleaving ribozymes that mediate processing of long multimeric transcripts into monomer-sized molecules in viroids, satellite RNAs, and retroviroid-like elements (for review, see Hammann and Lilley 2002). Moreover, both hairpin-like and hammerhead-like configurations are simple structural modules, very abundant in different functional RNA molecules, including aptamers, longer ribozymes, ribosomal RNAs, and viral and viroid RNA genomes (Wilson and Szostak 1999; Joyce 2004; Lilley 2005; Noller 2005). The two target structures used are represented in Figure 1.

Interacting populations

The scenario that we considered puts in a common environment the populations of independently selected modules, where they might covalently join to form molecules of double length. A cartoon of how this process is formally implemented in our numerical simulations is depicted in Figure 2. At each time step (defined by a replication cycle, as above), the two populations mix in a common pool. Whenever a correct hairpin structure is present, it catalyzes a number L of ligation reactions between two sequences from the same or different populations. L is a parameter of the model. The ligation is such that it joins the 3′ end of a randomly chosen sequence with the 5′ end of a second one. Those long sequences are folded and compared to the composed target structure (formed by the direct union of a hammerhead and a hairpin module). The simulation finishes at this point, since those long sequences do not replicate and are not subject to further selection. Finally, we assume that a short sequence folding into a hairpin structure cannot perform ligation catalysis if it is itself ligated to a second sequence.

In parallel, we consider a second scenario in which 70-nucleotides (nt)-long sequences (population C) evolve, following the same dynamical rules, toward the composed target structure, so that both functional modules must emerge simultaneously.

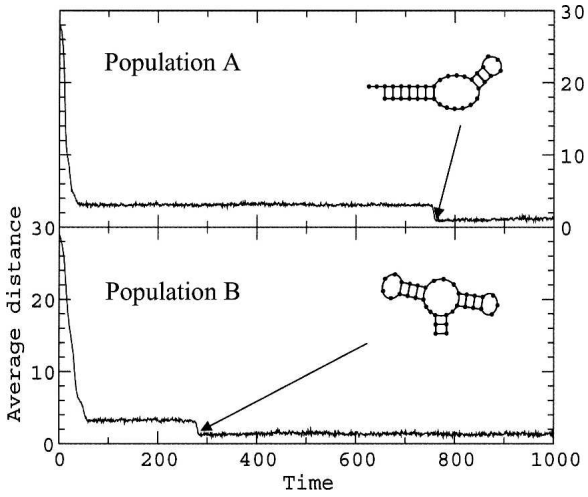

FIGURE 1. Two representative time series illustrating the dynamics of the RNA populations subject to replication with mutation and selection. The corresponding target structure (hairpin-like for population A and hammerhead-like for population B) is depicted in the *upper right* corner of each plot. For values of the mutation rate below the error threshold ($\mu = 0.005$ in these time series), a fraction of the population eventually folds into the target structure. The X axis represents time (number of generations). The Y axis represents the average distance of the secondary structures (corresponding to the sequences in the population) to the target structure. Once the mutation–selection equilibrium has been reached (indicated with arrows), the average distance of the population is not zero, since the relatively high mutation rate maintains a fraction of suboptimal mutants that necessarily coexist with the optimal class. In all cases, the population size is $N = 602$.

Computational results

Independent populations

Dynamics.

Time series illustrating the typical dynamics of populations A and B are represented in Figure 1. In these examples, evolution proceeds under a mutation rate $\mu = 0.005$ per

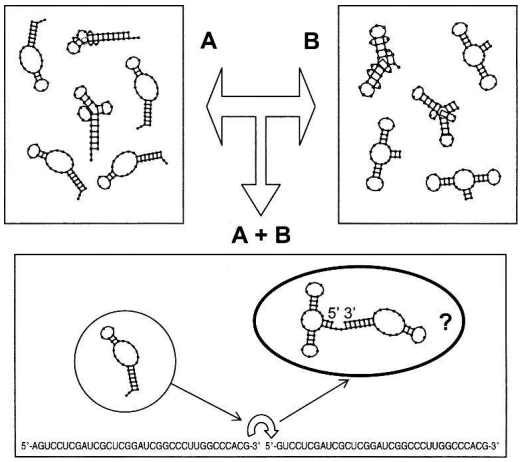

FIGURE 2. Schematic representation of the evolutionary model for molecular complexification. Two populations of 35-nt-long RNA sequences evolve independently with two target structures considered optimal in each environment: a hairpin-like structure (population A) and a hammerhead-like structure (population B). They attain statistically stationary selection–mutation equilibrium. At each time step, the two populations meet in a common pool and, whenever a hairpin structure is present, ligation reactions take place. In this context, we look for complex molecules whose minimum free-energy structure corresponds to the structure formed by directly joining the two small modules.

nucleotide and replication cycle. This mutation rate is relatively low for the current system, since it can maintain a large fraction of sequences folding into the target structure (∼90%) (see Fig. 3). The vertical axis of Figure 1 represents the average distance of the structures in the population to the target structure at each time step. For the parameters chosen, both populations approach very rapidly structures close to the objective. The arrow indicates the time at which the target structure is fixed, and the beginning of a statistically stationary state. For each value of μ, there is a fraction ρ of sequences yielding exactly the target structure at the stationary state. Whenever $\rho < 1$, structural variants are necessarily present in the population, and not all of the sequences fold into the target structure. The value of ρ for a fixed μ depends slightly on the structure chosen. Figure 3A represents the average value of ρ for the structures studied as a function of the mutation rate μ. We have also superimposed the values of ρ calculated for 25 independent realizations or replicas of the in silico experiment. The dispersion in this quantity is low, and thus most values fall close to the average ρ. As can be seen, for a large enough number of replication cycles and below a critical value $\mu_p \approx 0.039$, the target structure is always represented in the population. Above μ_p, the target structure cannot be fixed. That critical mutation rate corresponds to the phenotypic error threshold (Huynen et al. 1996; Takeuchi et al. 2005) above which it becomes impossible to faithfully transmit the minimal information required to maintain the structure of the molecules that constitute the quasispecies. The value of the phenotypic error threshold is strongly dependent on the length of the molecules in the population. For sequences of 70 nt, $\mu_p \approx 0.17$, meaning that there is an interval of mutation rates (between 0.017 and 0.039) where direct evolution of the full structure (as in population C) is not possible. The population wanders in the space of sequences without ever finding the objective structure.

Convergence time and neutral networks.

An important quantity in the evolutionary process is the time T_{Obj} required to reach the statistically stationary state, where the asymptotic density ρ stabilizes. The average value of T_{Obj} is represented in Figure 3B, as well as the values corresponding to 25 independent runs for each structure. The average time to attain each target structure has a minimum at values of the mutation rate large enough to steadily introduce novelty (in the form of new variants) in the population, but still sufficiently below the error

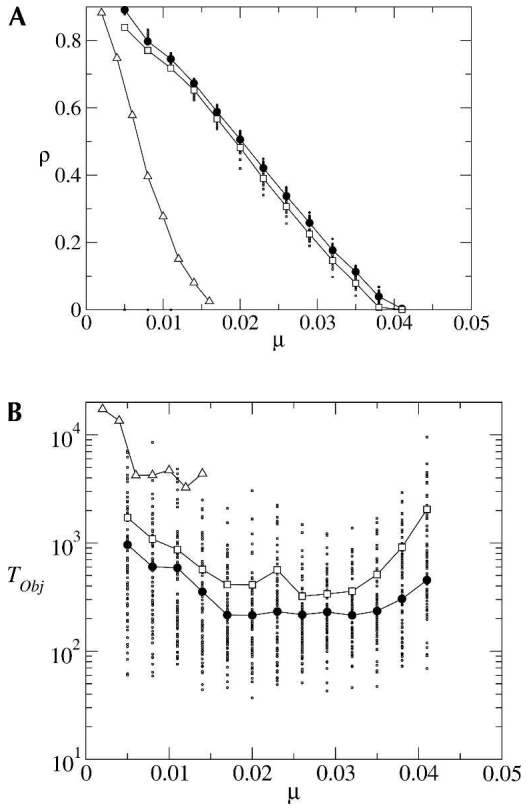

FIGURE 3. Stationary values of (*A*) the density ρ of sequences folding into the target structure and (*B*) the average number T_{Obj} of replication cycles required to attain the stationary state as a function of the mutation rate μ. Open squares correspond to the hairpin structure (population A), solid circles to the hammerhead (population B). Quantitative differences between the curves in either plot are due to differences in the structures used as objective. For comparison, the triangles represent ρ and T_{Obj} averaged over 25 runs for a population of sequences 70 nt in length (population C) that evolve toward the objective structure defined by the direct combination of the two small modules (depicted in Fig. 2).

threshold. There is thus an interval of μ values where the search process proceeds in an optimal way.

A second relevant observation is the following. While the density ρ is mostly independent of the realization (i.e., of the precise composition of the initial pool of sequences and of the random mutations occurring along the process), the time T_{Obj} varies broadly from run to run in each population. This reflects the extreme degeneracy of the sequence space when mapped to the structure space: for each run, the evolutionary process finds solutions to the required secondary structure through different pathways in the sequence space. Those pathways require broadly different generation times and lead to different domains in the neutral network: the master sequence of the quasispecies differs from realization to realization. The repetition of the same process and the comparison of the composition of the populations at the statistically stationary state is a way to probe the topological characteristics of the neutral network corresponding to each of the target structures. This comparison can be carried out by counting the difference in composition between pairs of sequences that have evolved within the same population (and thus have a common ancestor) and those resulting from two independent runs. While, within a population, the average distance between sequences is ∼4 nt for μ = 0.004 (see Fig. 4) and ∼7 nt for μ = 0.026 (data not shown), the average distance between populations with the same statistical characteristics and folding into the same secondary structure rises to 25–27 nt, and can be as large as 35 nt (two sequences folding into the same secondary structure differ in all of their nucleotides). This demonstrates that the neutral space of a structure typically percolates the sequence space. Such differences are of relevance when we consider composed structures, as will be seen.

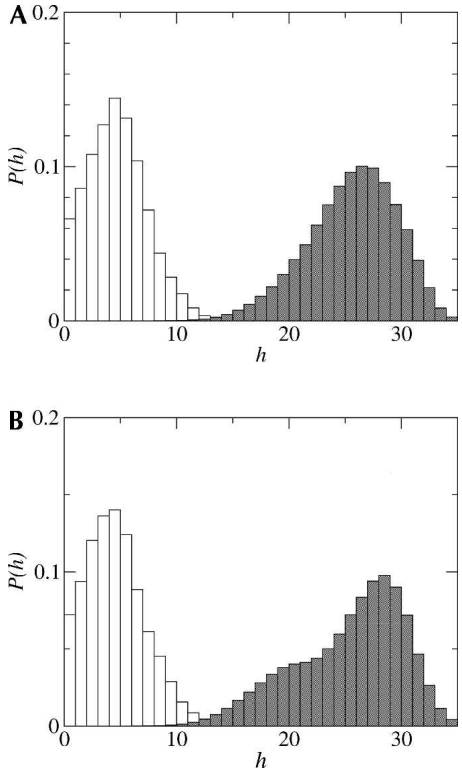

FIGURE 4. Histograms $P(h)$ of the difference in sequence composition h between pairs obtained in the same run and pairs corresponding to different runs. (*A*) Hairpin structure. The open histogram yields the probability that two sequences in the same population differ by h nucleotides. The shaded histogram yields the same difference for two sequences that have evolved in independent populations; (*B*) the same for the hammerhead structure. In both cases, the intrapopulation distances keep relatively low values, while the interpopulation distances can be as large as the length of the molecules. For μ = 0.004, which is the per nucleotide mutation rate used in those simulations, the density ρ of sequences folding into the target structure is ∼90% (see Fig. 3).

Formation of complex structures through ligation of modules

With the former results in mind, we investigate a scenario in which the two independent populations meet and experience ligation reactions, as explained above. Assuming that a sufficiently large number of functional hairpin structures is present, up to one-fourth of the ligation events will correctly join a sequence of population B (evolving toward a hammerhead motif) to a sequence of population A (evolving toward a hairpin). Note that the pairs that can be formed are A-A, A-B, B-A, and B-B. We call C_L the total number of ligation events that have occurred, irrespective of the original population of the sequences ligated. Out of that number, only B-A pairs have to be considered if we aim at obtaining a composed structure formed by a hammerhead plus a hairpin module (see Fig. 2). The number of pairs with the correct ordering is termed C_L^{B-A}.

All those B-A pairs are then folded. In principle, we expect a small number $F \ll C_L^{B-A}$ of those molecules to keep the independent modules in their folded configuration. Note that many new interactions between distant complementary nucleotides can arise in the ligated, long sequence, such that with a high probability the minimum free energy now corresponds to a secondary structure that differs from that maintaining the hammerhead and the hairpin as constitutive modules. Figure 5 represents the quantities C_L and F for 25 independent realizations of the process. While the total number of ligation events, C_L, is mainly independent of the realization, the number of folds maintaining the modules, F, varies widely. This reflects different degrees of compatibility between the composition of populations that have evolved independently, and the many solutions in sequence to the same structure, as shown in Figure 4.

Analytical results

We can gain some understanding of the interaction between populations A and B by estimating the expected

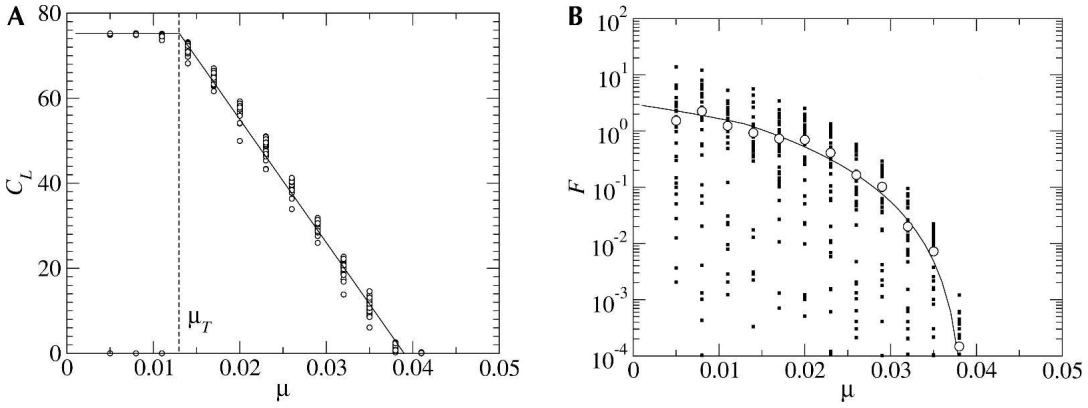

FIGURE 5. Result of the ligation interaction between the two populations folding into a hairpin and a hammerhead-like structure, respectively. (*A*) Number of successful ligation events per time step as a function of the mutation rate μ. Symbols are numerical results; the solid line is the analytical estimation, Equation 5. The vertical dashed line signals the separation between a regime where the catalyst is in excess ($\mu < \mu_T$) and a regime where it is in defect ($\mu > \mu_T$). For the parameters used ($N = 602$, $L = 1$), this transition occurs at $\mu_T = 0.013$. (*B*) Number of correctly folded complex structures per time step. Open circles correspond to averages over 25 independent realizations of the process; solid squares yield the result of each independent run. The solid line corresponds to the theoretical curve $F = (1/25)C_L\rho_H\,\rho_{HM}$ (Eq. 6).

number of ligation events C_L and the number of correct folds F in terms of the parameters of the system. When the two populations meet, there are $2N$ molecules of length n that can be ligated. Let us assume that out of those N_H correctly folds into the hairpin structure and catalyzes the ligation reaction. In our model, after a molecule acts as a ribozyme *in trans* (catalyzing a number L of reactions), it is degraded, such that it cannot be later ligated by another catalyst. Also, we do not allow ligation events performed by a hairpin motif once it is combined with a hammerhead one, even if the structure of the hairpin is correctly maintained in the complex molecule. The maximum number C_L of pairs that can be formed for $N_H > 0$ is thus

$$C_L = \text{int}\left[\frac{2N - N_H}{2}\right] \tag{1}$$

Let us first consider a situation in which the catalyst is in excess. This holds at least for low enough values of μ, where the density of hairpin-like structures has values close to 1. In this situation, N_H depends on L as

$$N_H = \frac{2N}{1+2L} \tag{2}$$

Now, there will be $N-N_H$ sequences corresponding to population A (hairpin) and N to population B (hammerhead), out of a total of $2N-N_H$ ready to form pairs. With these data, the number of correct C_L^{B-A} pairs of the form B-A is

$$C_L^{B-A} = \frac{N}{4}\left(1 - \frac{1}{2L}\right) \tag{3}$$

which approaches $(1/4)N$ in the limit of very large L, as expected.

The catalyst starts to be in defect at the point where the number of catalyzed reactions equals the number of correct pairs. The change of regime occurs at N_H such that $LN_H = C_L$. In order to proceed further, we need to make some assumptions on the number N_H. Looking at Figure 3A, where the density $\rho_H = N_H/N$ is represented, we see that a straight line with $\rho_H(0) = 1$ and $\rho_H(\mu_p) = 0$ fits very accurately the numerical results. [A linear dependence between the two quantities can be obtained in a first approximation within the context of quasispecies theory (Eigen 1971).] The transition from excess to defect of the catalyst takes place at a critical mutation rate

$$\mu_T = \mu_p \frac{2L-1}{2L+1} \tag{4}$$

Assuming thus that the density ρ decays linearly when μ increases, the amount C_L of successful ligation events also decays linearly between μ_T and μ_p, reaching zero at the latter value of the mutation rate. In this case, C_L becomes

$$C_L = \begin{cases} \dfrac{N}{4}\left(1 - \dfrac{1}{2L}\right) & \text{for} \quad \mu < \mu_T \\ \dfrac{N}{4}\left(1 - \dfrac{1}{4L}\right)\left(1 - \dfrac{\mu}{\mu_p}\right) & \text{for} \quad \mu_p > \mu > \mu_T. \end{cases} \tag{5}$$

This curve is represented in Figure 5A together with the numerical data. As can be seen, the agreement between numerical and theoretical results is quite good when we use the numerically estimated value of the phenotypic error threshold for the hairpin structure, $\mu_p \approx 0.039$.

Next, and using the results above, we estimate the fraction F of correct complex structures. The simplest

assumption is to consider that this fraction is proportional to the initial densities ρ_H and ρ_{HM} of correct hairpin and hammerhead structures, respectively, in the populations that meet. This leads to a straight forward estimation:

$$F = \alpha C_L \rho_H \rho_{HM} \quad (6)$$

The comparison between numerical and analytical results reveals that the theoretical curve fits quite well the average value of F for $\alpha = 1/25$, as can be seen in Figure 5B. Those analytical results are reliable only to a first approximation, since higher-order terms are needed in order to accurately represent the dependence between ρ and the mutation rate (Eigen and Schuster 1979; Schuster and Swetina 1988).

Modular scenario versus direct evolution of complex function

Modular evolution has clear quantitative advantages with respect to a direct evolution in which two or more functions (represented as molecular motifs) have to appear simultaneously in one single molecule. Whenever a structure corresponding to a sequence of length n can be divided into a number of modules, its appearance can be enhanced in the scenario here proposed. In this section, we derive some quantitative results in which the appearance of a structure corresponding to a sequence of 35 nt size is compared to the requirements needed for a secondary structure of 70 nt size to appear, as an example of the quantitative advantages offered by modular versus direct evolution.

Population size

It is possible to derive an analytic expression that relates the length of an RNA molecule to the (maximum) number of secondary structures that it can present. Schuster et al. (1994) used previous mathematical results (Stein and Waterman 1978) to estimate the number S_n of structures with hairpin loops of size three or more and without isolated base pairs (unstable structural elements), to find

$$S_n = 1.4848 n^{-3/2} (1.8488)^n \quad (7)$$

The number of sequences of length n is 4^n for the case of RNA, such that, assuming that all structures are equally likely, the probability P_n to obtain a given secondary structure from a random sequence of length n is

$$P_n = S_n / 4^n = 1.4848 n^{-3/2} (2.1636)^{-n} \quad (8)$$

[It is not true that all structures are probably equal. In fact, numerical analyses demonstrate that a small number of structures are "common." Asymptotically, almost all sequences fold into common structures (Grüner et al. 1996; Reidys et al. 1997). Hence, our estimation represents an upper bound for the population size required for "common" (and probably "natural") (see Gan et al. 2003) structures to be present.] The inverse of this number yields the minimum population size required for each structure to be present at least once. If the length of the sequences is 35 nt, a population of order 10^{14} molecules would contain roughly one representative of each structure. When the length is 70 nt, the required population size rises to 10^{26}. Actual experiments can easily proceed with molecular populations up to 10^{16}, while the sizes required to start with all representative structures of sequences of 70 nt are unattainable.

Evolutionary time

The possibility that the sought structure is already present in the initial population is of high relevance regarding evolutionary times, or number of replication cycles needed for the objective structure to be fixed in the population: once the structure is present, even in a small amount, the fixation time involves only a few replication cycles, and is largely independent of the length of the sequences in the population. The limiting step is thus to find the prefixed secondary structure when it is not present in the population. This is the case considered in our numerical simulations, since we assume that early molecular populations were small. Our preliminary results indicate that the search time T_{Obj} for the objective structure to be found grows exponentially with the length of the sequences in the population,

$$T_{Obj} \approx \exp(an) \quad (9)$$

While typical search times for $n = 35$ are of the order 10^2–10^3, this time increases 10-fold when the sequence length is doubled (see Fig. 3B). Simulations with sizes $n = 140$ are already infeasible, since they would require replication cycles of the order of 10^5 with the population sizes considered. We expect an exponential dependence between the two quantities also in experimental in vitro evolution, irrespective of the population size. Finally, let us mention that the time required for the two independent populations to ligate, a requisite for modular evolution, is negligible when compared to search and fixation times.

Mutation rates allowed

The interval of mutation rates where the objective structure can be attained through the evolutionary rules considered shrinks with the length of the sequences involved. This is an instance of the inverse dependence between the error threshold and the size of the evolving molecules. One example is shown in Figure 3, where it can be seen that the composed structure cannot be attained for mutation rates above 0.017. Thus, for values of μ between 0.017 and 0.039,

a structure of 70-nt length can only be obtained through modular evolution.

DISCUSSION

Small populations of replicating RNA sequences are able to evolve in a short time toward an optimal secondary structure. The length n of the sequences is bounded by the mutation rate, such that sequences are necessarily short if the copying accuracy is low. Selection on the secondary structure of the molecule (instead of selection on the precise sequence) yields a phenotypic error threshold μ_p larger than the genotypic error threshold, $\mu_p > \mu_g \approx 1/n$. The reason is that selection on the structure takes into account the degeneracy between the sequence and structure map (many different sequences yield the same secondary structure), and tolerance to mutations increases. This is an example of genetic or mutational robustness, that is, the preservation of a phenotype in the face of genetic perturbations (de Visser et al. 2003). Evolution of genetic robustness (where the selected sequence maximizes the number of neutral mutant neighbors) has been empirically detected, among others, in the genome of RNA viruses (Wagner and Stadler 1999) and in the structure of natural miRNAs (Borenstein and Ruppin 2006).

Under high mutation rates, however, constraints on the sequence length act against the appearance of complex, multifunctional molecules. This could have been the case of the first replicating polynucleotides at the early stages of the RNA world. A possible solution to this conundrum is the independent evolution of small functional modules or motifs that could later join to form longer, more complex molecules. We have shown that there are several advantages in this modular evolution of function. First, the probability that the large structure is directly found starting with correspondingly long sequences, and not as the result of combining independently evolved modules, is negligibly small. Second, the typical time required to select the small motifs is much shorter than that required to fix a structure in a population corresponding to a sequence double in length. Furthermore, in the 35-nt modules, the selection process is equally efficient (regarding the fraction ρ of sequences in the population folding into the target structure) at mutation rates double that of those allowed for sequences of 70-nt length (Fig. 3A). In addition to the smaller population sizes needed, the very possibility of finding complex structures, the shorter evolutionary times involved, and the allowance of larger mutation rates, are immediate advantages of the scenario involving modular evolution at early stages of the history of informative macromolecules.

Our results show that modular evolution can therefore fill a gap underlying the so-called molecular biologists' dream, at the point in chemical evolution where the first RNA molecules capable of copying themselves (acting as self-replicating RNA polymerase ribozymes) appeared (Szostak and Ellington 1993; Orgel 2004; Joyce and Orgel 2006). In our model, the starting material would be small populations of short RNA sequences coming from the nonenzymatic polymerization of nucleotides. It has been experimentally proven that montmorillonite-clay-catalyzed synthesis of RNA can render oligomers of length > 40 mers (Huang and Ferris 2003). Also, template-dependent polymerizations could have occurred in solution from oligonucleotides activated as phosphorimidazolides (Orgel 2004). Using those molecules as templates, highly mutagenic replication processes could have produced relatively large repertories of short, genetically different molecules, some of them folding into secondary/tertiary structures able to perform selectable functions. Different microenvironments could have promoted the selection of populations folding into different functional modules, which could then have mixed and ligated in progressively more complex molecules. Also, other functional capabilities could have appeared in two-domain molecules allowing long-range or allosteric effects, as well as a progressive increase in functional complexity.

In particular, modular evolution could have mediated the transition from an RNA ligase ribozyme into an RNA replicase. In vitro evolution experiments have shown that the simplest ribozyme that can catalyze RNA-dependent RNA polymerization is a complex, 189-nt-long RNA molecule composed of two relatively long structural modules (Johnston et al. 2001). Such in vitro evolution pathways, from a class I ligase ribozyme to an RNA polymerase ribozyme, as well as other approaches using different ligases as a starting point (for review, see Joyce 2004) are good examples of domain acquisition for the achievement of a new function. It is highly unlikely that such long RNA molecules (the shortest construct retaining RNA polymerase activity so far evolved is 165 nt long) (Johnston et al. 2001) could have originated through the high error rates assumed to have characterized initial replication mechanisms. Our model suggests that modular evolution could have progressively enriched short and simple RNA ligase ribozymes with additional RNA motifs, up to a total length of $\sim$200 nt. Such complex, more-sophisticated RNA enzymes could have acquired (among others) the ability to replicate any RNA template and, more importantly, to do it with progressively lower error rates. As it was recently stated: "competition for a limited supply of modules among a heterogeneous population of self-replicating ribozymes might provide the bases for Darwinian evolution" (Joyce 2004).

Apart from its possible implications for the origin of an RNA world, our modular evolution approach may shed light on the design and optimization of in vitro RNA evolution experiments. Such experiments generally proceed from a pool in the range of 10^{14}–10^{16} molecules ($\sim$1 nmol of material), which approximately corresponds to a combinatorial library containing one copy of all possible 25 mers

(Szostak and Ellington 1993). Nevertheless, in order to allow for sampling statistics, in vitro evolution should generally start with an average of 10–100 copies of each sequence (Joyce 2004). More importantly, the use of molecules only 25 nt in length may not permit the achievement of most of the catalytic functions, so that longer oligomers are required, with up to 200 random positions. Therefore, combinatorial libraries of long random sequences contain only a very reduced sample of all possible variant sequences (Wilson and Szostak 1999; Joyce 2004) and structures (see Results). This is a limiting factor of in vitro evolution technology, not only in terms of the time to attain the desired activity (rounds of stepwise evolution or transfers of continuous evolution), but also about the very possibility of reaching the required ribozyme or aptamer. Generally, this limitation is partially overcome by initiating the process with two or more synthetic random oligomers of shorter length that are amplified by PCR and ligated in order to construct the full-length population with which to start the in vitro evolution experiment (Szostak and Ellington 1993; Joyce 2004). Nevertheless, such a sequence-based approach may not significantly increase the overall representation of variants, limited by the (also negligible) fraction of short oligomers sampled.

Our results suggest, on the one hand, that a large increment in the potential for selecting complex structures or functions could be achieved through the design of independent in vitro evolution experiments of short modules (e.g., 20–40 nt in length) that are then allowed to ligate or recombine into longer molecules, and in turn are subjected to further evolution cycles or transfers. On the other hand, the initial steps involving short modules could be performed at a high mutation rate, therefore increasing the exploration of sequences that could result in an improved phenotype. Carefully designed in vitro evolution experiments should be carried out to test the validity of our in silico results. Computer simulations of structure-based modular evolution could complement in vitro evolution experiments in other ways as well. They can be useful in predicting or analyzing the results observed in experiments in which an RNA module is added to a preformed ribozyme or aptamer (Kumar and Joyce 2003; Romero-López et al. 2005), and in optimizing the effect of the combination of modules to develop allosteric ribozymes with relevant biotechnological applications (Komatsu et al. 2002; Penchovsky and Breaker 2005). Conversely, the structural invariability of the catalytic domain, which is susceptible to being computationally estimated, could be used as a key criterion in the design of experiments aimed at isolating the core elements of functional RNAs by trimming a longer, in vitro evolved aptamer or ribozyme (Lozupone et al. 2003; Wang and Unrau 2005). In summary, the modular evolution system presented here provides a framework for studying the increase of functional complexity in an RNA world. Furthermore, this approach can be helpful in the design of in vitro evolution experiments that seek improved functional molecules.

MATERIALS AND METHODS

RNA secondary structure is calculated using a simple and exact algorithm that takes into account the energetic contribution of pair formation, and discards the contribution due to loops of any kind (Nussinov et al. 1978; see also http://www.rpi.edu/zuckerm/MATH-4961/rnafold/node2.html). Although this algorithm is not accurate enough to yield precise structure predictions, it is fast and thus very convenient to undertake statistical analysis of population dynamics, which require extensive simulations. Furthermore, it is known that statistical properties of RNA secondary structure are almost independent of the algorithm used (Tacker et al. 1996), which guarantees the generality of the results derived in this framework.

The secondary structure is represented as a vector $S(i) = k$ of dimension n. The value of k either corresponds to the position of the nucleotide with which nucleotide i pairs or equals zero if the nucleotide belongs to a loop (i is unpaired). The secondary structure for each sequence is compared to a target structure $S^g(i)$, the latter representing an optimal function in the environment considered. The replicative ability of each sequence depends on its distance d to the target structure. The distance between structures corresponding to sequences of the same length is estimated as the ratio between the number of nucleotides that do not pair with the partner defined by the target structure and the total number of nucleotides:

$$d(S(i), S^g(i)) = \frac{\#mismatched\_pairs}{n} \tag{10}$$

In our model, the replicative ability of a sequence depends on the distance d. Specifically, the probability $p(S_j)$ that sequence j at replication cycle t is chosen as the parent sequence of a new sequence at replication cycle $t + 1$ is

$$p(S_j) = \frac{\exp\left(-\frac{d(S(j),\, S^g(j))}{\langle d\rangle_t}\right)}{\sum_{k=1}^{N} \exp\left(-\frac{d(S(k),\, S^g(k))}{\langle d\rangle_t}\right)} \tag{11}$$

where $\langle d\rangle_t$ is the average distance of the population to the target structure at time t.

The population is substituted by a new group of N sequences (resulting from their replication) at each time step. Every time that a sequence replicates, the daughter sequence mutates with probability μ per nucleotide. In this approach, no deletions or insertions are allowed during the replication, such that the length n of sequences is kept constant. The population size N is also constant throughout the simulations.

ACKNOWLEDGMENTS

We thank Drs. Ugo Bastolla, Alfredo Berzal-Herranz, and Michael Stich for valuable comments on the manuscript. This work was supported by Ministerio de Educación y Ciencia

(FIS2004-06414), INTA, EU, and CAM. S.C.M. benefits from a Ramón y Cajal contract.

Received June 22, 2006; accepted September 23, 2006.

REFERENCES

Ancel, L.W. and Fontana, W. 2000. Plasticity, evolvability, and modularity in RNA. *J. Exp. Zool. (Mol. Dev. Evol.)* **288:** 242–283.

Borenstein, E.B. and Ruppin, E. 2006. Direct evolution of genetic robustness in microRNA. *Proc. Natl. Acad. Sci.* **103:** 6593–6598.

Buzayan, J.M., Gerlach, W.L., and Bruening, G. 1986. Non-enzymatic cleavage and ligation of RNAs complementary to a plant virus satellite RNA. *Nature* **323:** 349–353.

de Visser, J.A.G.M., Hermisson, J., Wagner, G.P., Meyers, L.A., Bagheri-Chaichian, H., Blanchard, J.L., Chao, L., Cheverud, J.M., Elena, S.F., Fontana, W., et al. 2003. Perspective: Evolution and detection of genetic robustness. *Evolution Int. J. Org. Evolution* **57:** 1959–1972.

Doudna, J.A. and Cech, T.R. 2002. The chemical repertoire of natural ribozymes. *Nature* **418:** 222–228.

Eigen, M. 1971. Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58:** 465–523.

Eigen, M. and Schuster, P. 1979. *The hypercycle. A principle of natural self-organization*. Springer-Verlag, Berlin, Germany.

Fontana, W. 2002. Modelling ''evo-devo'' with RNA. *Bioessays* **24:** 1164–1177.

Fontana, W. and Schuster, P. 1998. Continuity in evolution: On the nature of transitions. *Science* **280:** 1451–1455.

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

Gilbert, W. 1986. Origin of life—The RNA world. *Nature* **319:** 618.

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

Hammann, C. and Lilley, D.M.J. 2002. Folding and activity of the hammerhead ribozyme. *Chembiochem.* **3:** 691–700.

Hampel, A. and Tritz, R. 1989. RNA catalytic properties of the minimum (−) sTRSV sequence. *Biochemistry* **28:** 4929–4933.

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

Huang, W. and Ferris, J.P. 2003. Synthesis of 35–40 mers of RNA oligomers from unblocked monomers. A simple approach to the RNA world. *Chem. Commun.* **12:** 1458–1459.

Huynen, M.A. 1996. Exploring phenotype space through neutral evolution. *J. Mol. Evol.* **43:** 165–169.

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

Jaeger, L., Wright, M.C., and Joyce, G.F. 1999. A complex ligase ribozyme evolved in vitro from a group I ribozyme domain. *Proc. Natl. Acad. Sci.* **96:** 14712–14717.

Johnston, W.K., Unrau, P.J., Lawrence, M.S., Glasner, M.E., and Bartel, D.P. 2001. RNA-catalyzed RNA polymerization: accurate and general RNA-templated prime extension. *Science* **292:** 1319–1325.

Joyce, G.F. 2002. The antiquity of RNA-based evolution. *Nature* **418:** 214–221.

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

Joyce, G.F. and Orgel, L.E. 2006. Progress towards understanding the origin of the RNA world. In *The RNA world*, 3rd ed. (eds. R.F. Gesteland et al.), pp. 23–56. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Knight, R. and Yarus, M. 2003. Finding specific RNA motifs: Function in a zeptomole world? *RNA* **9:** 218–230.

Komatsu, Y., Nobuoka, K., Karino-Abe, N., Matsuda, A., and Ohtsuka, E. 2002. In vitro selection of hairpin ribozymes activated with short oligonucleotides. *Biochemistry* **41:** 9090–9098.

Kumar, R.M. and Joyce, G.F. 2003. A modular, bifunctional RNA that integrates itself into a target RNA. *Proc. Natl. Acad. Sci.* **100:** 9738–9743.

Kun, Á., Santos, M., and Szahtmáry, E. 2005. Real ribozymes suggest a relaxed error threshold. *Nat. Genet.* **37:** 1008–1011.

Landweber, L.F. and Pokrovskaya, I.D. 1999. Emergence of a dual-catalytic RNA with metal-specific cleavage and ligase activities: The spandrels of RNA evolution. *Proc. Natl. Acad. Sci.* **96:** 173–178.

Lilley, D.M.J. 2005. Structure, folding, and mechanisms of ribozymes. *Curr. Opin. Struct. Biol.* **15:** 313–323.

Lozupone, C., Changayil, S., Majerfeld, I., and Yarus, M. 2003. Selection of the simplest RNA that binds isoleucine. *RNA* **9:** 1315–1322.

Luther, A., Brandsch, R., and von Kiedrowski, G. 1998. Surface-promoted replication and exponential amplification of DNA analogues. *Nature* **396:** 245–248.

Noller, H.F. 2005. RNA structure: Reading the ribosome. *Science* **309:** 1508–1514.

Nussinov, R., Piecznik, G., Grigg, J.R., and Kleitman, D.J. 1978. Algorithms for loop matchings. *SIAM J. Appl. Math.* **35:** 68–82.

Orgel, L.E. 2004. Prebiotic chemistry and the origin of the RNA world. *Crit. Rev. Biochem. Mol. Biol.* **39:** 99–123.

Penchovsky, R. and Breaker, R.R. 2005. Computational design and experimental validation of oligonucleotide-sensing allosteric ribozymes. *Nat. Biotechnol.* **23:** 1424–1433.

Prodi, G.A., Bakos, J.T., Buzayan, J.M., Schneider, I.R., and Bruening, G. 1986. Autolytic processing of dimeric plant virus satellite RNA. *Science* **231:** 1577–1580.

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

Reidys, C., Stadler, P.F., and Schuster, P. 1997. Generic properties of combinatory maps: Neutral networks of RNA secondary structures. *Bull. Math. Biol.* **59:** 339–397.

Robertson, M.P. and Ellington, A.D. 2000. Design and optimisation of effector-activated ribozyme ligases. *Nucleic Acids Res.* **28:** 1751–1759.

Romero-López, C., Barroso-del Jesús, A., Puerta-Fernández, E., and Berzal-Herranz, A. 2005. Interfering with hepatitis C virus IRES activity using RNA molecules identified by a novel in vitro selection method. *Biol. Chem.* **386:** 183–190.

Schultes, E.A. and Bartel, D.P. 2000. One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science* **289:** 448–452.

Schuster, P. 1993. RNA based evolutionary optimization. *Orig. Life Evol. Biosph.* **23:** 373–391.

Schuster, P. and Swetina, J. 1988. Stationary mutant distributions and evolutionary optimization. *Bull. Math. Biol.* **50:** 635–660.

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

Stein, P.R. and Waterman, M.S. 1978. On some new sequences generalizing the Catalan and Motzkin numbers. *Discrete Math.* **26:** 261–272.

Szostak, J.W. and Ellington, A.D. 1993. In vitro selection of functional RNA sequences. In *The RNA world*, 1st ed. (eds. R.F. Gesteland and J.F. Atkins), pp. 511–533. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

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

Takeuchi, N., Poorthuis, P.H., and Hogeweg, P. 2005. Phenotypic error threshold; additivity and epistasis in RNA evolution. *BMC Evol. Biol.* **5:** 9.

Tang, J. and Breaker, R.R. 1997. Rational design of allosteric ribozymes. *Chem. Biol.* **4:** 453–459.

van Nimwegen, E., Crutchfield, J.P., and Huynen, M.A. 1999. Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci.* **96:** 9716–9720.

Vlassov, A.V., Johnston, B.H., Landweber, L.F., and Kazakov, S. 2004. Ligation activity of fragmented ribozymes in frozen solution: Implications for the RNA world. *Nucleic Acids Res.* **32:** 2966–2974.

Wagner, A. and Stadler, P.F. 1999. Viral RNA and evolved mutational robustness. *J. Exp. Zool. (Mol. Dev. Evol.)* **285:** 119–127.

Wang, Q.S. and Unrau, P.J. 2005. Ribozyme motif structure mapped using random recombination and selection. *RNA* **11:** 404–411.

Westhof, E. and Massire, C. 2004. Evolution of RNA architecture. *Science* **306:** 62–63.

Wilke, C.O. 2001. Selection for fitness vs. selection for robustness in RNA secondary structure folding. *Evolution Int. J. Org. Evolution* **55:** 2412–2420.

Wilson, D.S. and Szostak, J.W. 1999. In vitro selection of functional nucleic acids. *Annu. Rev. Biochem.* **68:** 611–647.