Chapter 4
Populations of RNA Molecules as Computational Model for Evolution

Michael Stich, Carlos Briones, Ester Lázaro, and Susanna C. Manrubia

Abstract We consider populations of RNA molecules as computational model for molecular evolution. Based on a large body of previous work, we review some recent results. In the first place, we study the sequence–structure map, its implications on the structural repertoire of a pool of random RNA sequences and its relevance for the RNA world hypothesis of the origin of life. In a scenario where template replication is possible, we discuss the internal organization of evolving populations and its relationship with robustness and adaptability. Finally, we explore how the effect of the mutation rate on fitness changes depends on the degree of adaptation of an RNA population.

4.1 Introduction

Molecular evolution covers a huge area of research, ranging from prebiotic chemistry and questions on the origin of life, through many aspects related to the origin of and the relationships among species, the study of viral and bacterial evolution and their medical implications up to the artificial design and in vitro selection of molecules, with all their applications in nano- and biotechnology. In this chapter, we do not aim to give a complete overview of that wide research field, but focus on the use of populations of RNA molecules as a model to understand evolution of prebiotic replicators in the RNA world. As RNA viruses share many characteristics with primitive RNA molecules with replicative ability, these studies can also be used to tackle many aspects of viral evolution. Although a large body of our work is inspired by experiments, in this chapter we focus on theoretical approaches for understanding evolutionary processes.

M. Stich, C. Briones, E. Lázaro, and S.C. Manrubia
Dpto de Evolución Molecular, Centro de Astrobiología (CSIC-INTA), Ctra de Ajalvir, km 4, 28850 Torrejón de Ardoz (Madrid), Spain
e-mail: stichm@inta.es

P. Pontarotti (ed.), *Evolutionary Biology – Concepts, Molecular and Morphological Evolution*, DOI 10.1007/978-3-642-12340-5_4,

RNA molecules are a very well suited model for studying evolution 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 (phenotype) of the molecule. Since in many cases the spatial structure of the molecule is crucial for its biochemical function, the structure of an RNA molecule can be considered as a minimal representation of the phenotype.

In current biology, RNA viruses are the paradigmatic example for evolving populations: replication is fast, it takes place with a relatively high error rate, and population sizes are large. This has made RNA viruses an often used example for *quasispecies*, a concept originally proposed by Eigen (1971) and developed over the last decades in the context of virology (Domingo 2006). It states that a population of replicators, e.g., an RNA virus evolving within an infected host, cannot be represented by only one, fittest, genome, but by the spectrum of related mutants that are present in the population. The quasispecies evolves under a certain error (mutation) rate and the cloud of mutants enables the population to adapt quickly to new environmental situations, such as population bottlenecks and changed selective pressures. Under constant external conditions, a quasispecies approaches a dynamic equilibrium between selection of favorable sequences (what we mean by favorable, will be specified below) and the diversity constantly introduced by mutation. Therefore, the mutation rate is of crucial importance in the study of such heterogeneous populations in molecular evolution (Huynen et al. 1996; Biebricher and Eigen 2005): if the mutation rate becomes too large, selection becomes inefficient, the correlations between the genomes within the population decay, and the whole population may even become extinct. There are many reported examples of the extinction of RNA virus populations when replication takes place at increased error rates due to the presence of mutagenic agents (Sierra et al. 2000; Domingo 2005; Cases-González et al. 2008). These results have inspired a new promising antiviral strategy named lethal mutagenesis (Loeb et al. 1999).

Another field of research within molecular evolution is the quest for understanding the origin and early evolution of life. One of the most appealing theories in this context is the so-called RNA world hypothesis. It is based on the facts that RNA cannot only represent a genetic code, like DNA in present-day cells, but also can act as catalyst of biochemical reactions, like present-day enzymes. Therefore, a single RNA molecule could have been endowed with the two main features of living matter, providing the genome (i.e., the blueprint for replication) and the primordial machinery for replication and metabolism. One of the open questions in this context is how the first template-dependent RNA polymerase ribozyme could have emerged. Experimentally, a minimum size of approximately 165 nucleotides has been established for such a molecule (Johnston et al. 1999; Joyce 2004), a length three to four times that of the longest RNA oligomers obtained by random polymerization (Huang and Ferris 2003, 2006). Hence, one of the main challenges within the RNA world scenario is to convincingly bridge this gap.

In this chapter, we will review some recent results obtained in our lab (Manrubia and Briones 2007; Stich et al. 2007, 2008, 2010; Briones et al. 2009) and put them into the context of the aforementioned issues. The first part of this chapter tries to deepen our understanding of the sequence–structure map, relevant for the RNA world model. Then, we discuss the internal organization of evolving populations and its relevance for robustness and adaptability. Subsequently, we explore the relationship between microscopic mutation rate and the fractions of beneficial and deleterious mutations, as observed in experiments or used in phenomenological models.

4.2 Structural Repertoire of RNA Pools

RNA structure is crucial for biochemical function of an RNA molecule. A lot of research efforts are dedicated to the folding process that relates RNA sequences with RNA structures. For our purpose, it is sufficient to consider two-dimensional secondary structures as good approximation of real three-dimensional structures. Two fundamental properties of the sequence–structure map are that (1) the number of different sequences is much higher than the number of structures and (2) not all possible structures are equally probable (Fontana et al. 1993; Schuster et al. 1994). In this context, *common* structures are those which have many different sequences folding into them and *rare* structures are those which have only few sequences folding into them. In this section, we explore the structural repertoire of a pool of random sequences.

We first describe the results of the folding of 10^8 RNA molecules of length 35 nt consisting of random sequences composed of the four types of nucleotides A, C, G, and U (Stich et al. 2008). As secondary structure of each molecule, we take the minimum free energy structure as given by the `fold` () routine from the Vienna RNA Package (Hofacker et al. 1994).

RNA secondary structures consist of stems, where base pairing (A–U, G–C, G–U) between nucleotides occurs, and unpaired regions. In standard bracket notation, nucleotides paired with each other are denoted by "(" and ")", while unpaired nucleotides are represented by ".". Among unpaired regions, we can distinguish dangling ends and different kinds of loops: hairpin loops, bulges, interior loops, and multiloops. The simplest structure is called a stem–loop, it consists of one hairpin loop and one stem, and possibly one or two dangling ends. While there are 4^n sequences of length n (the so-called 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 given (Grüner et al. 1996). Therefore, different sequences will actually fold into the same secondary structure, grouping into neutral networks of genomes (Grüner et al. 1996; Huynen et al. 1996). Neutral networks are formed by genomes sharing the same phenotype, here secondary structure, and which are connected by (single) mutational events. The sequence–structure map turns out to be very complex. Two sequences that are just one mutation apart may fold into structures very different from each other. At the same time, in a relatively small neighborhood of any sequence, almost all common structures can be found (Fontana et al. 1993).

In our case, 10^8 sequences folded into 5,163,324 structures (Stich et al. 2008). A way to visualize the uneven distribution of sequences into structures is the frequency–rank diagram. In Fig. 4.1a, we have ranked the structures according to the number of sequences folding into them. One can see that there are around thousand common structures, each of them obtained from about 10^4 different sequences. On the other hand, we also find a few million rare structures yielded by only one or two sequences. Although for a much smaller pool, this has already been reported before (Schuster et al. 1994; Grüner et al. 1996; Schuster and Stadler 1994; Tacker et al. 1996).

In order to study the distribution of common vs. rare structures in more detail, we have proposed a classification where we characterize a structure in terms of three numbers (Stich et al. 2008): (a) the number of hairpin loops, H, (b) the sum of bulges and interior loops, I, and (c) the number of multiloops, M. For example, a simple stem–loop structure, denoted as SL, is characterized by (H,I,M) = (1,0,0), and all stem–loop structures found in the pool are grouped into that *structure family*. Other important families are the hairpin structure family, HP, with one interior loop or bulge (1,1,0), the double stem–loop, DSL, represented by (2,0,0), and the simple hammerhead structure, HH, by (2,0,1). Of course, there exist more complicated structure families, as detailed in Stich et al. (2008). For the pool that we have folded, we find that only 21 structure families are enough to cover all the 5.2 million structures identified.

Our analysis, displayed in Fig. 4.1b, shows that the vast majority of sequences fold into simple structure families. For example, 79.0% of all sequences belong to only three structure families (HP, HP2, SL, in decreasing abundance), and 92.1% of all sequences fold into simple structures with at most 3 stems (HP, HP2, SL, DSL, DSL2, HH). Note that 2.1% of all sequences remain open and do not fold. Our data is in agreement with other findings on the structural repertoire of RNA sequence

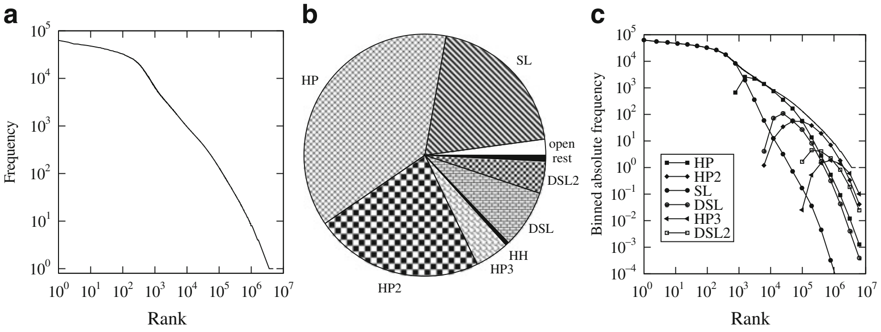

Fig. 4.1 (**a**) Frequency–rank diagram of the 5,163,324 different secondary structures, obtained by folding 10^8 RNA sequences of length 35 nt. (**b**) Distribution of the sequences in structure families according to their frequency. Higher-order hairpins, HPx, are defined as (H,I,M) = (1,x,0), being $x \geq 2$, higher-order double stem–loops, DSLx, as (H,I,M) = (2, $x-1$,0), and higher-order hammerheads, HHx, as (2, $x-1$,1). (**c**) Frequency–rank diagram according to the structural family. The upper thick solid curve denotes the same curve as in (**a**). Parts (**a**) and (**c**) after Stich et al. (2008)

pools where the influence of the sequence length (Sabeti et al. 1997; Gevertz et al. 2005), the nucleotide composition (Knight et al. 2005; Kim et al. 2007), and pool size (Gevertz et al. 2005) has been studied.

Now, we can reconsider the frequency–rank diagram. We sum up all structures of a given structure family within a rank interval. Through this binning procedure, we obtain for each structure family a curve which describes its relative frequency compared with that of the other families. The curves for the most frequent families are shown in Fig. 4.1c. We immediately see that the most frequent structures belong to the stem–loop family, followed by the hairpin family, double stem loops, higher-order hairpin families, and hammerheads. For low ranks, the SL curve is identical with the curve describing all structures. For ranks between 4×10^3 and 10^4, it is the HP curve which practically coincides with the total curve. Interestingly, the position of the bump around rank 10^3 falls together with the locations where the SL and HP families are equally present. Hence, we conclude that the bumps in the frequency–rank diagram correspond to the succession of different structural families and are not smoothed by better sampling of the sequence space.

What implications have these findings for the RNA world scenario? The standard view of the RNA world hypothesis states that the first chains of polymerized polynucleotides consisted of random sequences. Therefore, it is important to study the structural and subsequently the functional repertoire of such short sequences. We have seen that a random pool is very rich in simple structures. However, as already mentioned above, short molecules cannot perform template-dependent replication. Therefore, we devised a four-step model of modular evolution as a possible pathway for the emergence of functional and progressively longer molecules starting with a random pool of RNA oligomers (Briones et al. 2009). The first step is the random polymerization of RNA molecules up to 40-mers. The second step is the folding of these sequences, leading to high fractions of simple structures like hairpins, as just shown. The third step is based on the observation that simple hairpin structures, similar to those formed by short random sequences in huge amounts, are actually known to show catalytic activity, leading to RNA–RNA ligation (Puerta-Fernández et al. 2003). If a certain fraction of the hairpin molecules originated is capable of displaying ligase activity, longer molecules may be formed. Even though the majority of the long molecules may not perform ligase activity, some of them will keep the modular structure of their building blocks and remain active to catalyze further RNA–RNA ligations (Manrubia and Briones 2007). This suggests that hairpin ribozymes, both in individual modules and in combined structures, could have catalyzed the synthesis of progressively longer RNA molecules from short and structurally simpler modules (Briones et al. 2009). Finally, the fourth step of the model consists of a maturation of these ligating RNA molecules of intermediate length into self-replicating RNA ligase networks, which could coexist and even compete with each other, leading eventually to a molecule long and complex enough to perform template-dependent RNA replication [further details in Briones et al. (2009)]. It is important to emphasize that the whole model relies strongly on the observation that simple structures like hairpins – with potential ligase activity – are ubiquitous in pools of random RNA sequences.

4.3 Internal Organization of Evolving Populations

Above, we have discussed the static picture of the sequence–structure map. Once replication within a population is possible, evolution through Darwinian selection is triggered. Here, RNA serves as a model to study the interplay between mutation, selection, and the diversity sustained in populations of fast mutating replicators (Stich et al. 2007).

First, we briefly describe the evolutionary algorithm. Our system consists of a population of N replicating RNA sequences, each of length n nucleotides. At the beginning of the simulation, every molecule is initialized with a random sequence. Every time that a sequence replicates, each of its nucleotides has a probability μ (mutation rate) to be replaced by another nucleotide, randomly chosen among the four possibilities A, C, G, U.

At each generation, the sequences are folded into secondary structures as described above. We define a target structure that represents in a simple way optimal performance in a given environment. It can be a hairpin, hammerhead, or any other structure: the qualitative behavior of the system does not depend on this choice. We compare every folded structure 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 a given structure into the target structure (Hofacker et al. 1994). The closer a secondary structure is to the target structure, the higher the probability $p(d_i)$ that the corresponding sequence i replicates:

$$p(d_i) = \frac{\exp(-\beta d_i)}{\sum_{i=1}^{N} \exp(-\beta d_i)}. \tag{4.1}$$

The parameter β denotes the selective pressure and is here chosen as $\beta = 2/n$. Generations in our simulations are nonoverlapping and the offspring generation is calculated according to Wright–Fisher sampling.

Two relevant quantities to characterize the state of the population are the average distance $d = \sum_{i=1}^{N} d_i/N$ to the target structure and the fraction ρ of structures in the population folding exactly into the target structure. Because of the persisting action of mutation, both quantities fluctuate in time even after reaching the asymptotic regime. Therefore, we perform averages over long time intervals (and different realizations, starting from distinct initial RNA populations), obtaining mean values denoted by $\bar{d}$ and $\bar{\rho}$, respectively.

In order to quantify collective properties of the molecular ensemble, we first determine the consensus sequence of the population, given by, for each position along the sequence, the most frequent type of nucleotide found within the population. In real RNA molecular and viral quasispecies, the consensus sequence is obtained by means of population sequencing (Thurner et al. 2004; Simmonds et al. 2004; Domingo 2006), and it does not necessarily correspond to any of the individual sequences present in the population. It is straightforward to fold the consensus sequence and obtain the *structure of the consensus sequence*, for which its coincidence with the target structure can be determined. At each time step we count either one, corresponding to coincidence, or zero, otherwise. Averages over time (and realizations) of this binary variable yield $\bar{\rho}_{\mathrm{C}}$, which corresponds to the probability that, at a randomly chosen time step, the structure of the consensus sequence coincides with the target structure.

We further define a *consensus structure*. It is calculated by determining, for each position along the molecule, the most frequent structural state found within the population, i.e., unpaired ".", paired upstream "(", or paired down-stream ")". Due to this definition, the consensus structure does not necessarily represent a valid secondary structure of an RNA molecule. This procedure is hence fundamentally different from assigning a consensus structure to an alignment of sequences (Hofacker et al. 2002). Averages over time (and realizations) of the coincidence between the consensus structure and the target structure yield the probability $\bar{\rho}_{\mathrm{S}}$.

Within this model, evolution takes place in the following way: sequences which fold into structures similar to the target structure will replicate more likely and their fraction in the population increases. Mutation introduces diversity and enables the system to find structures that are closer to the target, and finally find and fix the target structure. Starting from a random set of sequences, we can distinguish several phases of evolution: the search phase, where d decreases while $\rho = 0$. This phase finishes at generation g_{A} when a molecule folds into the target structure for the first time. Then, the phase of fixation begins, where – on average – d still decreases and ρ increases. However, due to the stochastic nature of mutation – and hence in particular for large mutation rates as will be explored further below – the population may lose again the target structure (and ρ drops down to zero). If ρ does not drop to zero for 500 consecutive generations, we say that the target structure has been fixed at generation g_{F}. Then, the asymptotic regime is reached, where d and ρ fluctuate around constant values and which corresponds to a mutation–selection equilibrium. If the mutation rate μ is too large, the population is unable to maintain the target structure within the population. In absence of an analytic theory for the system we are studying, we determine the fixation threshold as the value μ_{F} at which the curve $g_{\mathrm{F}}(\mu)$ diverges.

Since we now have defined the main quantities to describe the population, we show the results in Fig. 4.2. They were obtained from simulations for a system of $N = 1{,}000$ RNA molecules of length $n = 30$ nt evolving toward a hairpin structure. In (a) we show the curves for $\bar{\rho}, \bar{\rho}_{\mathrm{C}}$, and $\bar{\rho}_{\mathrm{S}}$. The quantity $\bar{\rho}$ describes the fundamental property of a quasispecies at mutation–selection equilibrium. For small μ, $\bar{\rho}$ takes maximal values. This means that a population contains the largest fraction of correctly folded molecules if it evolves at small mutation rates. As μ increases, $\bar{\rho}$ decreases monotonously until it approaches zero. To determine the fixation threshold, we look at Fig. 4.2b where we show the curves of the search time and search plus fixation time. The solid curve represents the search time. We observe that for small μ finding the target structure is difficult because only little diversity is introduced and the search process is slow. Therefore, fixation takes a long time. As μ increases, the introduced diversity in the population becomes larger and both search and search plus fixation times decrease. However, fixation turns out to be a

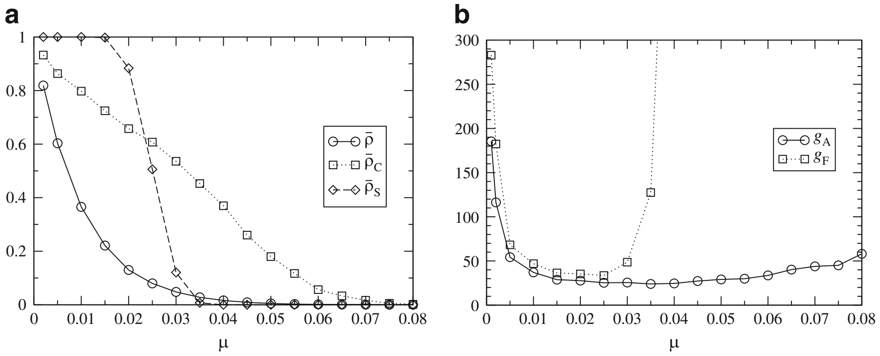

Fig. 4.2 (**a**) Asymptotic properties of a population of size $N = 1{,}000$ and molecules of length $n = 30$ nt as function of the mutation rate μ. Displayed are the average fraction of correctly folded structures $\bar{\rho}$, and the quantities $\bar{\rho}_{\mathrm{C}}$ and $\bar{\rho}_{\mathrm{S}}$. Averaging has been performed over 4,000 generations and 20 realizations, disregarding the first 2,000 generations. (**b**) Search time g_{A} and search plus fixation time g_{F}. We locate the fixation threshold where g_{F} diverges. Averaging has been performed over 200 realizations. The population evolves toward a hairpin target structure given by ..((((((. . .(((. . .))). . . .)))))) in bracket notation

difficult task if μ is too large, and the curves for search and search plus fixation start to deviate. The search plus fixation time g_{F} (dotted curve) diverges around $\mu \approx 0.045$, where we approximately locate the fixation threshold for this n and target structure. This means that while the population shows largest $\bar{\rho}$ for small μ and highest degree of diversity close to the fixation threshold, the search and fixation times are optimized for intermediate mutation rates around $\mu \approx 0.025$ well below the fixation threshold.

Coming back to Fig. 4.2a, we now have a look at the curves for $\bar{\rho}_{\mathrm{C}}$ and $\bar{\rho}_{\mathrm{S}}$. The curve of $\bar{\rho}_{\mathrm{C}}$ lies for all considered mutation rates above the curve of $\bar{\rho}$. This means that based upon the information of the consensus sequence only, one may overestimate the evolutionary success. This effect is observed both below and above the fixation threshold. For example, for $\mu = 0.05$, where only 0.5% sequences fold into the target structure, and only into an intermittent way, the probability that the consensus sequence folds into the target structure is still 18%. Consequently, the population remains close to sequences that actually fold into the target structure although it is unable to fix it. Obviously, this is related to the fact that at least part of the population are descendents from the same sequence and hence are closely related to each other. Note that the probability that a sequence of the population folds into the target structure is different from the probability that the consensus sequence does. Since consensus sequences are readily obtained from molecular or viral quasispecies, one should take into account this difference.

Considering now the curve for $\bar{\rho}_{\mathrm{S}}$, we observe a qualitatively different behavior: for $\mu < 0.025$, the probability that the consensus structure coincides with the target structure is practically one, while for $\mu > 0.025$, it approaches zero. For small μ, this effect can be easily explained: the weight of all the correctly folded molecules is strong enough to keep $\bar{\rho}_{\mathrm{S}}$ high. But in Stich et al. (2007), we showed that even neglecting the correctly folded molecules and for large mutation rates, among the remaining sequences there is a sufficiently large fraction of those molecules which have a similar structure to the target structure. An analogous effect is known for random sequences: in a small neighborhood of a given sequence, the most probable structures are identical or very similar to the structure of the reference sequence (Fontana et al. 1993). Even where $\rho_S = 0$, the distribution of the structure states along the chain may still resemble the target structure and the positions where the concordance is broken correspond to positions that are actually less stable.

While $\bar{\rho}_C$ senses the similarity among the sequences and $\bar{\rho}_S$ the similarity among the structures, both quantities take superior values than $\bar{\rho}$ for most of the mutation rates in spite of the fact that selection is actually acting upon structure (not sequence) and that the corresponding fitness landscape is rough. This means that the population retains relevant structural information in a distributed fashion even above the fixation threshold. This represents a strong structural robustness and suggests that certain functional RNA secondary structures may effectively withstand high mutation rates (Stich et al. 2007).

4.4 Phenotypic Effect of Mutations

In the last section, we have already discussed the optimal mutation rate to promote adaptation in an evolving system. Here, we calculate the distribution of the effects of mutations on fitness and the relative fractions of beneficial and deleterious mutations (Stich et al. 2010). It is important to recall that the effect of mutations on the phenotype depends on the genomic and populational context. We explore two different situations: the mutation–selection equilibrium (equilibrated population) and the first stages of the adaptation process (adapting population).

Here, we consider a population of $N = 1{,}000$ molecules of length $n = 50$ nt evolving toward a hairpin target structure. The change in fitness of an RNA sequence under replication is quantified by the change of distance to the target structure, i.e., by $\Delta_{ij} = d_i - d_j$, where i denotes the mother and j the daughter sequence. Hence, for $\Delta_{ij} > 0$ ($\Delta_{ij} < 0$), the mutations lead to an increase (decrease) of fitness and hence are beneficial (deleterious). If $\Delta_{ij} = 0$, either no mutation occurred or the mutations had no effect on fitness (were neutral). As we sum up over N values of Δ_{ij} at each generation (and over generations and realizations as specified below), we obtain a probability distribution $\Pi(\Delta)$ of the changes in fitness.

In Fig. 4.3a, we show for three different mutation rates the distributions $\Pi(\Delta)$, obtained for populations at mutation–selection equilibrium. The part of the distribution with the largest weight represents replication events with no or neutral mutations ($\Delta = 0$). For a very low mutation rate, negative fitness events strongly dominate over the positive ones and hence beneficial mutations are rare. As the mutation rate increases, the curves move up for positive and negative Δ since there are more mutation events. Although in particular beneficial mutations occur more often, negative fitness effects still dominate in absolute numbers.

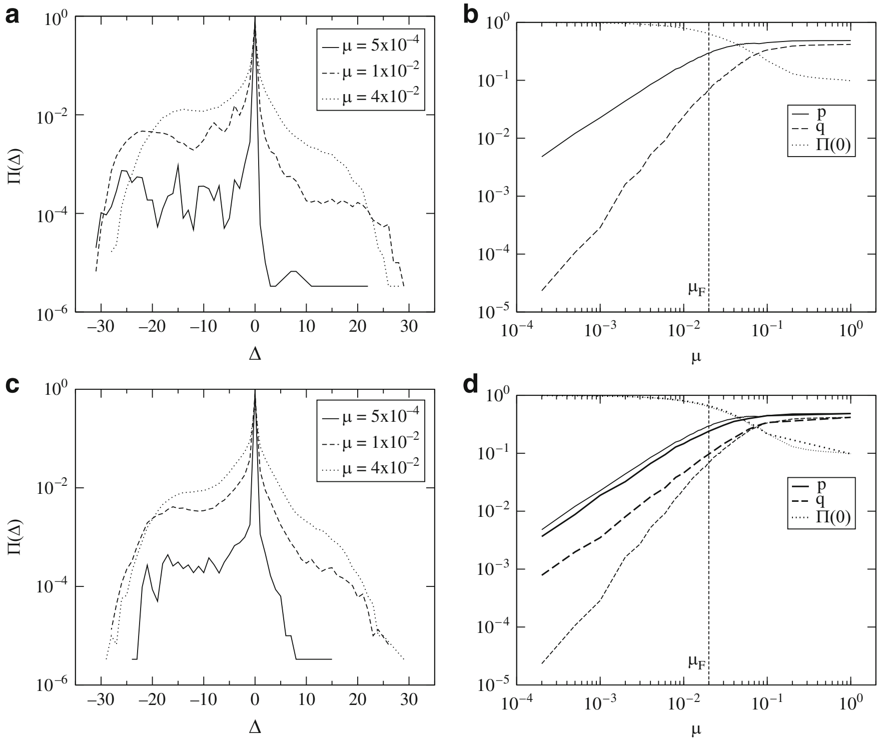

Fig. 4.3 Phenotypic changes of mutations for optimized (**a**, **b**) and adapting (**c**, **d**) populations. (**a**) Probability distribution $\Pi(\Delta)$ obtained from 300 generations in the asymptotic regime and for three different values of μ. (**b**) Beneficial (q) and deleterious (p) *phenotypic* mutation rates as function of the *microscopic* mutation rate μ for optimized populations. Replication events without fitness change are given by $\Pi(0)$. (**c**) As (**a**), but for adapting populations (probability distributions obtained from the first 50 generations and 6 different realizations). (**d**) As (**b**), but for adapting populations. The thin curves denote the curves from (**b**). The target structure is ((((. . . . .(((((. (((((. . . . . .))))).))))). . . . .)))) in bracket notation. After Stich et al. (2010)

From the distribution Π we can calculate the fraction of deleterious changes p and beneficial changes q in the following way:

$$q = \int_{0+}^{\infty} \Pi(\Delta)\mathrm{d}\Delta, \tag{4.2}$$

$$p = \int_{-\infty}^{0-} \Pi(\Delta)\mathrm{d}\Delta. \tag{4.3}$$

These quantities represent the beneficial and deleterious *phenotypic* mutation rates which shall not be confounded with the *microscopic* mutation rate μ. By definition, $p + q + \Pi(0) = 1$.

How q and p depend on μ is depicted in Fig. 4.3b. For low mutation rates, we see that p is more than two orders of magnitude larger than q. As μ increases, both p and q increase, although $p > q$ for all μ, in particular for mutation rates below the fixation threshold, for this n and target structure approximately located at $\mu_{\mathrm{F}} = 0.02$. As μ increases, the fraction of replication events with no change in fitness, given by $\Pi(0)$, decreases. The ratio p/q decreases from more than two orders of magnitude to less than one close to μ_{F}. This reflects the fact that the higher the mutation rate at which a population has reached mutation–selection equilibrium the lower the fraction of correctly folded molecules, and hence beneficial mutations are more probable. However, these beneficial mutations do not increase the degree of adaptation of the population due to the difficulties to get fixed at high error rate.

In Fig. 4.3c,d, we show the distribution $\Pi(\Delta)$ and the functional behavior of $(p, q) = f(\mu)$ for adapting populations. In this case, fitness changes are measured before the target structure has been found. The distributions $\Pi(\Delta)$ behave in a qualitatively similar way, although quantitative differences to Fig 4.3a can be seen, e.g., for $\mu = 0.0005$: The range of negative Δ is smaller than for an equilibrated population, so very deleterious mutations are not present, and also the overall level of deleterious mutations is lower. At the same time, beneficial mutations are more common. This observation can be explained by the fact that since the population is still relatively far from target, mutations that drive a sequence even further are less likely. For the same reason, mutations that have a positive effect on fitness are more probable.

Figure 4.3d summarizes the results: In an adapting population, p is smaller than at equilibrium, and q is larger, although these differences get much lower as the error rate increases. However, in all cases there are still more deleterious mutations than beneficial ones. Again, both phenotypic mutation rates increase as μ increases, while replication events without phenotypic change decrease.

4.5 Summary

Here, we have presented recent results with RNA populations as computational model to explore and understand evolutionary processes, using the complex underlying sequence–structure–function relationship of RNA molecules.

In the first section, we showed some observations on the structural repertoire of random RNA sequences (Stich et al. 2008). One important result is that simple structures like stem–loops and hairpins are dominant in pools of short sequences. This finding, together with other results and arguments, allowed us to devise a stepwise model of modular evolution for the origin of the RNA world (Briones et al. 2009).

In the second section, we introduced an algorithm of RNA evolution in silico (Stich et al. 2007). After characterizing the asymptotic state of the population (at mutation–selection equilibrium), we showed that search and fixation times are optimized for intermediate mutation rates, far from the fixation threshold where the creation of diversity is maximal and far from the regime of low mutation rates where evolutionary success is optimized (in terms of correctly folded molecules). These results have important implications for the adaptability of virus and replicator populations that, due to the changes in the selective pressures that they continuously experience, need to have the capability to adapt rapidly, which can be obtained by the selection of high mutation rates. However, the difficulties for the fixation of beneficial mutations, together with the low fitness values attained when replication takes place at mutation rates close to the error threshold, suggest that viral quasispecies operate at mutation rates considerably smaller.

Furthermore, close to and even beyond the fixation threshold, RNA populations show clear signatures of the target structure they try to approach (Stich et al. 2007). For example, even a population that contains practically no molecule that folds into the correct structure, as a whole may actually harbor the target structure as the structure of its consensus sequence. This demonstrates that the evolutionary success of the population is more robust than suggested by the spectrum of its mutants alone.

Finally, we have established a connection between the microscopic mutation rate μ and the phenotypic mutation rates p and q (Stich et al. 2010). These mutation rates are used in phenomenological models of population dynamics and also in fitting models of data obtained from experiments (Eyre-Walker and Keightley 2007). We find that adapting populations have a much larger fraction of beneficial mutations than equilibrated ones, especially for small mutation rates. Furthermore, we have shown that increases in μ do not cause linearly proportional increases in p and q, as often assumed in simple models of population evolution.

In summary, our results encourage the combined approach of experimental research and computational modeling for studying molecular evolution.

Acknowledgments The authors acknowledge support from Spanish MICIIN through projects FIS2008-05273 and BIO2007-67523, from INTA, and from Comunidad Autónoma de Madrid, project MODELICO (S2009/ESP-1691).

References

Biebricher CK, Eigen M (2005) The error threshold. Virus Res 107:117–127

Briones C, Stich M, Manrubia SC (2009) The dawn of the RNA world: Toward functional complexity through ligation of random RNA oligomers. RNA 15:743–749

Cases-González C, Arribas M, Domingo E, Lázaro E (2008) Beneficial effects of population bottlenecks in an RNA virus evolving at increases error rate. J Mol Biol 384:1120–1129

Domingo E (ed) (2005) Virus entry into error catastrophe as a new antiviral strategy. Virus Res 107:115–228

Domingo E (ed) (2006) Quasispecies: concept and implications for virology. Springer, Berlin

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

Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. Nat Rev Genet 8:610–618

Fontana W, Konings DAM, Stadler PF, Schuster P (1993) Statistics of RNA secondary structures. Biopolymers 33:1389–1404

Gevertz J, Gan HH, 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 IL, Stadler PF, Schuster P (1996) Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks. Monatsh Chem 127:355–374

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P (1994) Fast folding and comparison of RNA secondary structures. Monatsh Chem 125:167–188

Hofacker IL, Fekete M, Stadler PF (2002) Secondary structure prediction for aligned RNA sequences. J Mol Biol 319:1059–1066

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

Huang W, Ferris JP (2006) One-step, regioselective synthesis of up to 50-mers of RNA oligomers by montmorillonite catalysis. J Am Chem Soc 128:8914–8919

Huynen MA, Stadler PF, Fontana W (1996) Smoothness within ruggedness: the role of neutrality in adaptation. Proc Natl Acad Sci USA 93:397–401

Johnston WK, Unrau PJ, Lawrence MS, Glasner ME, Bartel DP (1999) RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. Science 292:1319–1325

Joyce GF (2004) Directed evolution of nucleic acid enzymes. Annu Rev Biochem 73:791–836

Kim N, Gan HH, 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

Loeb LA, Essigmann JM, Kazazi F, Zhang J, Rose KD, Mullins JI (1999) Lethal mutagenesis of HIV with mutagenic nucleoside analogs. Proc Natl Acad Sci USA 96:1492–1497

Manrubia SC, Briones C (2007) Modular evolution and increase of functional complexity in replicating RNA molecules. RNA 13:97–107

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

Sabeti PC, Unrau PJ, Bartel DP (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, Stadler PF (1994) Landscapes: complex optimization problems and biopolymer structures. Comput Chem 18:295–324

Schuster P, Fontana W, Stadler PF, Hofacker IL (1994) From sequences to shapes and back: a case study in RNA secondary structures. Proc R Soc Lond B Biol Sci 255:279–284

Sierra S, Dávila M, Lowenstein PR, Domingo E (2000) Response of foot-and-mouth disease virus to increased mutagenesis. J Virol 74:8316–8323

Simmonds P, Tuplin A, Evans DJ (2004) Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: implication for virus evolution and host persistence. RNA 10:1337–1351

Stich M, Briones C, Manrubia SC (2007) Collective properties of evolving molecular quasispecies. BMC Evol Biol 7:110

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

Stich M, Lázaro E, Manrubia SC (2010) Phenotypic effect of mutations in evolving populations of RNA molecules. BMC Evol Biol 10:46

Tacker M, Stadler PF, Bornberg-Bauer EG, Hofacker IL, Schuster P (1996) Algorithm independent properties of RNA secondary structure predictions. Eur Biophys J 25:115–130

Thurner C, Witwer C, Hofacker IL, Stadler PF (2004) Conserved RNA secondary structures in *flaviviridae* genomes. J Gen Virol 85:1113–1124

Waterman MS (1978) Secondary Structure of Single-stranded Nucleic Acids. In: Rota G-C (ed) Studies in Foundation and Combinatorics, vol 1 of: Advances in Mathematics Supplementary Studies. Academic Press, New York, pp 167–212