Evolution of RNA-Based Networks

Peter F. Stadler

Abstract RNA molecules have served for decades as a paradigmatic example of molecular evolution that is tractable both in in vitro experiments and in detailed computer simulation. The adaptation of RNA sequences to external selection pressures is well studied and well understood. The de novo innovation or optimization of RNA aptamers and riboswitches in SELEX experiments serves as a case in point. Likewise, fitness landscapes building upon the efficiently computable RNA secondary structures have been a key toward understanding realistic fitness landscapes. Much less is known, however, on models in which multiple RNAs interact with each other, thus actively influencing the selection pressures acting on them. From a computational perspective, RNA–RNA interactions can be dealt with by same basic methods as the folding of a single RNA molecule, although many details become more complicated. RNA–RNA interactions are frequently employed in cellular regulation networks, e.g., as miRNA bases mRNA silencing or in the modulation of bacterial mRNAs by small, often highly structured sRNAs. In this chapter, we summarize the key features of networks of replicators. We highlight the differences between quasispecies-like models describing templates copied by an external replicase and hypercycle similar to autocatalytic replicators. Two aspects are of importance: the dynamics of selection within a population, usually described by conventional dynamical systems, and the evolution of replicating species in the space of chemical types. Product inhibition plays a key role in modulating selection dynamics from survival of the fittest to extinction of unfittest. The sequence evolution of replicators is rather well understood as approximate optimization in a fitness landscape for templates that is shaped by the sequence-structure map of

P.F. Stadler (✉)
Institute Für Informatik der Universität Leipzig, Härtelstraße 16-18,
04107 Leipzig, Germany
e-mail: peter.stadler@bioinf.uni-leipzig.de

P.F. Stadler
Max Planck Institute for Mathematics in the Sciences, Inselstraße 22,
04103 Leipzig, Germany

P.F. Stadler
The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Current Topics in Microbiology and Immunology (2016) 392: 43–59
DOI 10.1007/82_2015_470

Published Online: 04 September 2015

RNA. Some of the properties of this map, in particular shape space covering and extensive neutral networks, give rise to evolutionary patterns such as drift-like motion in sequence space, akin to the behavior of RNA quasispecies. In contrast, very little is known about the influence of sequence-structure maps on autocatalytic replication systems.

Contents

1 From Replication to Fitness Landscapes

1.1 The Quasispecies

The interplay of replication, selection, and mutation is the basis of Darwinian evolution. Replication can be interpreted as an autocatalytic chemical reaction that, in the simplest case, is of the form

$$\mathbf{A} + \mathbf{Y} \xrightarrow{k} 2\mathbf{Y}, \tag{1}$$

where **A** is the substrate that is used as building material for the autocatalyst **Y** that is required as template for its own formation. Mutation in the form of imprecise, or error-prone, reproduction represents the universal kind of variation, which occurs in all organisms and can be sketched by a single overall reaction step:

$$\mathbf{A} + \mathbf{Y} \xrightarrow{k_{xy}} \mathbf{Y} + \mathbf{X}. \tag{2}$$

Here, the mutant is denoted by **X,** and the rate parameters k and k_{xy} refer to two parallel reaction channels. Reaction rates depend explicitly on the *type* y of the replicator **Y**. These rate differences are the basis for selection due to the fact that the different templates compete for the common resource **A**.

In the case of RNA or protein molecules, y is simply the nucleotide sequence of the molecule $\mathbf{Y}$. The production of a variant $\mathbf{X}$ from $\mathbf{Y}$, in the simplest model, occurs with a probability $q_{xy} := \mathbb{P}[\mathbf{Y} \to \mathbf{X}]$, while $\mathbf{Y}$ undergoes the copying reaction with a rate $k_0(\mathbf{Y}) = \alpha f(y)$, where $\alpha := [\mathbf{A}]$ is the concentration of the building material $\mathbf{A}$. Adding an unspecific outflow from the system, we arrive at a

$$\frac{\mathrm{d}[x]}{\mathrm{d}t} = \sum_y q_{xy} \alpha f(y)[y] - \varphi[x] \tag{3}$$

The dynamics of this system is thus described by two ingredients: the rate of copying of each type, $f(x)$, usually referred to as its *fitness*, and the probabilities of specific copying errors q_{xy}. The specific form of the flux φ plays little role for the overall dynamics as long as it is small enough not to completely drain all the replicating material from the system (Happel and Stadler 1999). In the simplest case, we may assume that the total concentration of all replicating types as well as the concentration of the building material is kept constant. This model, known as *constant organization*, yields the famous quasispecies equation as follows:

$$\frac{\mathrm{d}[x]}{\mathrm{d}t} = \sum_y q_{xy} f(y)[y] - \varphi[x] \quad \text{with} \quad \varphi = \sum_{x,y} q_{xy} f(y)[y] \tag{4}$$

i.e., φ equals the total production of replicating types.

Since the molecular types x and y are sequences, the most common model assumes that $\begin{pmatrix} q_{xy} = \\ ndp^d \end{pmatrix}$ where n is the sequence length, $d = d_{xy}$ is the number of sequence positions in which x and y differ, n is the common length of both sequences, and p is the probability of a point mutation, i.e., the exchange of a single letter, per copying event. Nearly all work on the quasispecies model has used this mutation model (Eigen 1971; Eigen et al. 1989; Eigen and Schuster 1977). Conceptually, it can be simplified further by ignoring the small probabilities of multiple mutations, simply setting $q_{xy} = 0$ if $x \neq y$ differ by more than a single mutation. Sequences are "easily accessible," i.e., adjacent, if they differ by a single (point) mutation only. This arranges the set X of sequences as the vertices of a graph, which is usually referred to as the *sequence space* (X, τ). The symbol τ denotes the accessibility structure, which here is just the edge set of graph but may, in general, be a more complicated topological structure (Flamm et al. 2007). Together with the fitness function $f : X \to \mathbb{R}$, that sequence space forms the *fitness landscape* (X, τ, f).

The dynamics of a population evolving according to Eq. (4) is governed by the underlying fitness landscape. Its stationary solution

$$\sum_y q_{xy} f(y)[y] = \varphi[x] \tag{5}$$

determines the equilibrium distribution of the variants in an evolved population. In the limit of small mutation rates, where $q_{xy} \approx 1 - pa_{xy}$ with $a_{xy} = 1$, if x and y are adjacent and $p \ll 1$, one can show that $[x]$ is concentrated around the fitness maximum (Eigen 1971). Chapters 1, 4, and 5 in this book are concerned with the relationships of the fitness landscape (X, τ, f) and the structure of the resulting quasispecies.

1.2 Molecular Replicators

A variety of in vitro systems embody self-replication of RNA or DNA. Although template-instructed ligation can also be achieved without enzymes, these are restricted to short and usually specialized sequences (Lee et al. 1996; von Kiedrowski 1986). All copy reactions of interestingly long and diverse templates known today, however, require elaborate enzymes.

The earliest system studied in detail was based on the RNA-dependent RNA polymerase of the bacteriophage Qβ, a Levivirus. Extensive studies on the reaction kinetics of this system (Biebricher and Eigen 1988) demonstrated that the kinetic data are consistent with a many-step reaction mechanism describing the stepwise addition of nucleotides. It can be coarse-grained to a Michaelis–Menten-like overall reaction of the form

$$\mathbf{A} + \mathbf{Y} + \mathbf{E} \leftrightharpoons \mathbf{A} + \mathbf{Y} \cdot \mathbf{E} \rightarrow \mathbf{Y} \cdot \mathbf{E} \cdot \mathbf{X} \leftrightharpoons \mathbf{Y} \cdot \mathbf{E} + \mathbf{X} \tag{6}$$

that still explains the observed three distinct regimes: exponential growth a low $[\mathbf{Y}]$, linear growth for intermediate replicator concentrations, and saturation by product inhibition at high concentrations. The rate constants depend strongly on the sequence of the template $\mathbf{Y}$ since Qβ replicase is well adapted to recognize the genomic RNA of the Qβ phage and to discriminate it from host cell sequences. Affinity to the replicase thus is an important determinant of fitness in in vitro evolution experiments with this system. Manipulation of the environment in this experimental setup has lead to the selection of widely different RNA molecules with surprising properties. SV11, for instance, is replicated from an extremely stable metastable conformation of the RNA (Biebricher and Luce 1992); a “drug-addicted” RNA was obtained in Kramer et al. (1974) by adding the intercalating dye ethidium bromide.

Enzymes that replicate nucleic acid templates effectively independent of their sequence have evolved in particular for genomic DNA templates. The discovery of the DNA polymerase chain reaction (PCR) (Mullis et al. 1986) was a milestone toward sequence independent amplification of DNA sequences in vitro. It requires, however, higher temperatures to separate the two strands of the double helical product. Since the product of template-directed replication or ligation is invariably a double strand, product inhibition cannot be entirely avoided in most systems. The simplest replication system is thus better described by

$$\mathbf{A} + \mathbf{Y} + \mathbf{E} \rightarrow \mathbf{E} + \mathbf{Y} \cdot \mathbf{Y} \quad \text{and} \quad \mathbf{Y} \cdot \mathbf{Y} \leftrightharpoons 2\mathbf{Y} \tag{7}$$

Under the "quasi-steady state" approximation (Segel and Slemrod 1989), this model follows a modified kinetics (Wills et al. 1998)

$$\frac{\mathrm{d}[x]}{\mathrm{d}t} = [x](f(x)\psi(\beta(x)[x]c) - \varphi) \tag{8}$$

where $f(x)$ is the fitness measures at infinite dilution, i.e., for a total concentration of replicating material $c \rightarrow 0$, $\beta(x)$ is a constant derivable from the microscopic rate constants that describes the strength of product inhibition, and $\psi(u) = 2(\sqrt{u+1} - 1)/u$ is a function that decays like the square root of its argument. It reduces to the simple case for $\beta(x) \rightarrow 0$. A wide range of related mechanisms of template-directed ligation, including an experimentally studied systems based on DNA triple helices (Li and Nicolaou 1994), and the membrane-anchored mechanism suggested for the "Los Alamos Bug" artificial protocell project (Rasmussen et al. 2003) follow the same effective kinetic law. An approximation that replaces $\psi(u)$ by $\sqrt{u}$ was considered in Szathmáry and Gladkih (1989).

In contrast to the quasispecies-like models, which (apart from the mutant cloud around the "master sequence") effectively lead to "survival of the fittest" at least for small mutation rates, there is no selection in the parabolic growth model (Varga and Szathmáry 1997; Wills et al. 1998). More general systems with product inhibition, however, allow for cooperation of all replicators whose fitness exceeds a certain concentration-dependent threshold, which can be computed explicitly (Wills et al. 1998). Most of the experimental systems of self-replicating polymers without enzyme fall into this class, see, e.g., Plöger and Kiedrowski (2014) for a recent peptide nucleic acid system. Unless product inhibition is too strong (or concentrations become too large), these systems show selection by "extinction of the unfittest" rather than survival of the fittest. So far, their population-level dynamics have not been explored for complex, realistic fitness landscapes.

Several enzymatic systems have been well established to amplify nucleic acid sequences, most famously the PCR (Erlich 1989). Here the product inhibition problem is solved by "thermal cycling," i.e., a periodic increase in temperature to release product bound in duplices. An isothermal version based on T7 polymerase is the 3SR reaction (Fahy et al. 1991). A recently introduced alternative is the isothermal multiple displacement amplification (IMDA) (Luthra and Medeiros 2004). A common theme in these technologically highly relevant systems is that templates are amplified nearly independently of their sequence. This enables among other applications high-throughput DNA sequencing as well as artificial selection (Systematic Evolution of Ligands by EXponential Enrichment (SELEX)) (Ellington and Szostak 1990; Tuerk and Gold 1990). In the latter, amplification is alternated by an assay that enriches RNA or DNA molecules with desired properties, thereby effectively implementing a user-defined fitness landscape. This allowed the creation of RNA and DNA sequences with a surprisingly wide variety of both binding and enzymatic properties. A detailed mathematical analysis of the SELEX procedure and its convergence properties can be found in Levine and Nilsen-Hamilton (2007).

1.3 Quantifying Natural Fitness Landscapes

Since the volume of sequence space increases exponentially with sequence length, it has long been impossible to obtain a comprehensive picture of fitness landscapes. Nevertheless, early attempts to empirically chart at least a neighborhood of the optimal or native sequence date back almost two decades (Aita et al. 2002; Aita et al. 2000; Hayashi et al. 2006; Reetz and Sanchis 2008).

The situation has changed with the advent of micro-arrays and then deep sequencing technologies. At least for very small model systems, it has become possible to measure the relative abundance of a very large number of sequences. Comparing the sequence distribution $p(x)$ with the distribution in the initial pool $p_0(x)$ immediately yields the estimated

$$f(x) \approx p(x)/p_0(x)\,. \tag{9}$$

It is worth noting that the fitness function cannot be inferred from the equilibrium quasispecies distribution since the latter is usually concentrated around the fittest member of the population, making it impossible to obtain data for more distant parts of the landscape.

Micro-arrays have provided a convenient means of measuring the fitness of larger samples in parallel (Lauring and Andino 2011). Instead of measuring fitness directly from an adapting population, equilibrium parameters such as RNA-protein binding constants also have been measured using micro-arrays (Rowe et al. 2010). Various sources of bias deriving from the ligation and sequencing steps must be measured and taken into account in the practical analysis of HTS-based surveys of landscapes, see, e.g., Jimenez et al. (2013) for details. Earlier work still targeted particular regions in sequence space. The fitness of all possible individual point mutants of a nine-amino acid region of yeast Hsp90 was determined in Hietpas et al. (2011). Pitt and Ferré d’Amaré (2010) mutagenized an artificial RNA ligase ribozyme and estimated the fitness in the neighborhood of the original ligase sequence from a reselected variant pool.

Various interpolation and machine learning schemes have been proposed to estimate the structure of fitness landscapes from sparse data (Romero et al. 2013; Woo and Reifman 2014). The issue, however, remains a difficult one and the problem is far from being solved. Most geometric and topological characteristics of the inferred fitness landscape, such as the number of monotonically increasing paths between ancestral and derived genotypes, the prevalence of sign epistasis, and the number of local fitness maxima, are distorted in the inferred landscape (Otwinowski and Plotkin 2014).

Most recently, small systems have been investigated in their entirety. An example is a survey the space of 24-mers selected for GTP binding (Jimenez et al. 2013), see also Athavale et al. (2014) for a recent review.

Empirical landscapes seem to differ substantially between different biological systems. DNA-protein binding landscapes (Rowe et al. 2010) have been found to be rugged, with many local optima. Protein landscape, on the other hand, seems to be at least locally smooth, "Mt. Fuji-like" (Aita et al. 2000; Lobkovsky et al. 2011) consistent with computational prediction from simple computational models (Babajide et al. 1997; Chan and Bornberg-Bauer 2002). Similar structures are observed for viral fitness landscapes, both experimentally (Lauring and Andino 2011; Woo and Reifman 2014) and in computational models (Kouyos et al. 2012). For RNA landscapes, however, there is an apparent discrepancy between the observed very rugged landscapes and the computationally predicted landscape structure (Athavale et al. 2014). We return to this issue in some detail in the next section.

A promising alternative to exhaustively covering sequence space is to associate fitness with a low-dimensional phenotype. Acevedo et al. (2014), for instance, mapped thousands of measured fitness values of mutants onto three-dimensional structures of viral proteins to explore the structure–function relationships.

1.4 Computer Models of RNA Evolution

RNA has turned out to be a particularly fruitful model to study fitness landscapes in computational models of evolution. On the one hand, it is intriguing catalytic activities and its crucial involvement in the core information metabolism of modern life forms, having lead up to the RNA world hypothesis of pre-biotic evolution (Gilbert 1986), justify detailed studies. On the other hand, nucleic acid structures have specific feature, not shared by polypeptides, which make them computationally tractable at a convenient coarse-grained resolution.

Nucleobases interact specifically through hydrogen bonding in a manner that established simple combinatorial rules of complementary base pairing (GC, AU, and GU). Paired bases form regular helical structures stabilized by π-electron interactions whose energetics are nearly perfectly additive in terms of the contributions of adjacent base pairs. The resulting secondary structures shown in Fig. 1 are thus *matchings* in the graph theoretical sense, which are further restricted by the *noncrossing* rule that excludes so-called "pseudoknots." These simplifications result in the combinatorial model of RNA secondary structures in which the *folding problem*, i.e., the prediction of structure from sequence, can be solved efficiently and completely by means of dynamic programming (Zuker and Stiegler 1981). This simple model, which has been parameterized by careful thermodynamic measurements (Turner and Mathews 2010), has proved its practical relevance in thousands of applications ranging from explaining and organizing structural features of RNAs to the prediction of effects of mutations and wholesale design of functional RNAs. Far from perfect, it nevertheless captures most of the energetics of RNA folding, it

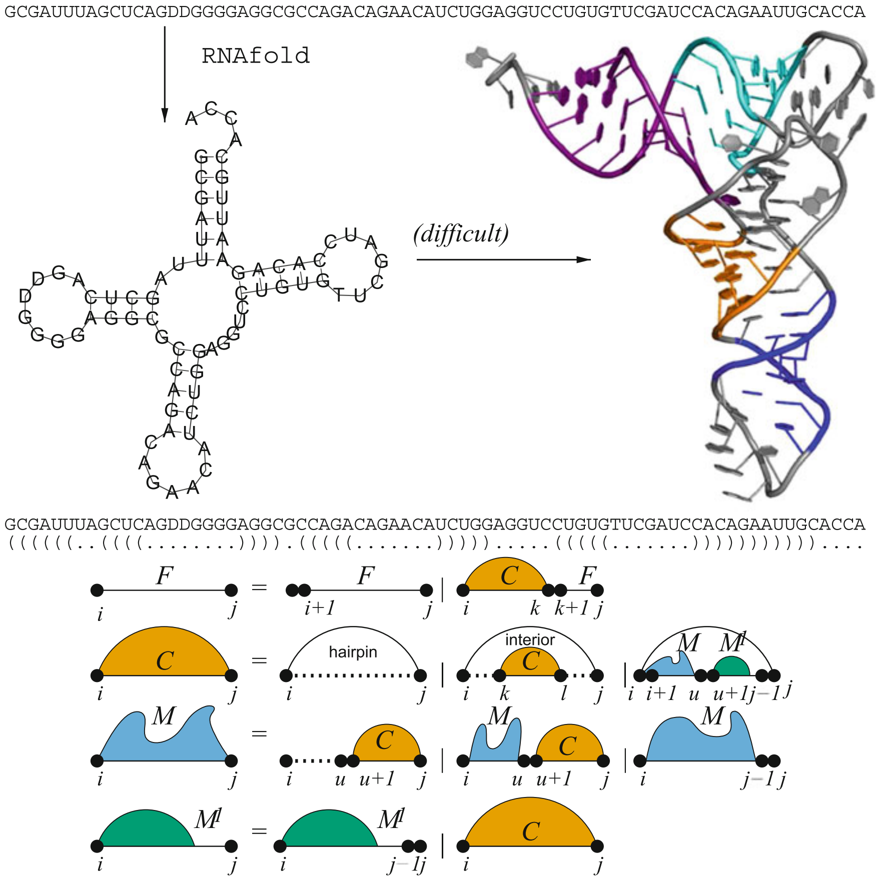
GCGAUUUAGCUCAGDDGGGGAGGCGCCAGACAGAACAUCUGGAGGUCCUGUGTUCGAUCCACAGAAUUGCACCA
RNAfold
(difficult)
GCGAUUUAGCUCAGDDGGGGAGGCGCCAGACAGAACAUCUGGAGGUCCUGUGTUCGAUCCACAGAAUUGCACCA
((((((..((((........)))).(((((.......))))).....(((((.......)))))))))))....
F
i
j
=
i+1
C
k
k+1
hairpin
interior
l
M
M1
u
u+1
j−1

Fig. 1 RNA folding in a nutshell. *Top* Folding of the yeast tRNA-Phe. The secondary structure can be computed with little effort, while the 3D structure (shown here is the PDB crystal structure 5TNA) is not easily accessible by computational methods alone. *Below* The algorithm of RNA folding consists of fairly simple recursion relations that construct the energy (or partition function) of a substructure on a sequence interval from i to j from smaller components. An arbitrary structure (F) begins with either an unpaired base • or a substructure enclosed by a base pair (C). In both cases, it then continues with a correspondingly shorter unconstrained structure. The *second line* describes the decomposition of base pair enclosed structure (C) into the three major loop types: hairpin loops, interior loops (including stacked base pairs, $k = i + 1$ and $l = j - 1$), and multi-branch loops. The last three lines correspond to the recursion for multi-branch loops, see Lorenz et al. (2011); Zuker and Stiegler (1981) for details. For each possible decomposition step, the energy of the *l.h.s.* structure is the sum of the energies of the *r.h.s.* components. These recursions require quadratic memory and cubic time in terms of the input sequence length, providing a highly efficient and exact solution of the RNA folding problem

describes key features of RNA folding kinetics, and it explains many of the evolutionary patterns observed in RNA. We refer to a recent book on RNA bioinformatics for details on applications and limits of the model (Gorodkin and Ruzzo 2014).

Instead we concentrate here on a different aspect which was explored in substantial detail almost a quarter of a century ago (Schuster et al. 1994). The efficiency of structure prediction has made it feasible to explore the sequence-structure map of RNA as a proxy for genotype–phenotype maps. Starting from the insight that genotype (sequence) is acted upon by mutation and other genetic operators while the phenotype (structure) is subject to selection, it is appealing to model biologically relevant fitness landscapes as compositions

$$f(x) = \phi(\Phi(x)) \tag{10}$$

where $\Phi : X \to \mathbb{P}$ is the genotype–phenotype map and $\phi : \mathbb{P} \to \mathbb{R}$ is a fitness function that evaluates the phenotypes $y \in \mathbb{P}$ rather than the genotype. In the case of RNA secondary structures, Φ is simply RNA folding as implemented, e.g., by the `ViennaRNA` package (Lorenz et al. 2011) and $\mathbb{P}$ denotes the set of RNA secondary structures. In many circumstances, the properties of $f : X \to \mathbb{R}$ are essentially determined already by the genotype–phenotype map Φ, as in the case of RNA secondary structures.

Extensive computational studies (Fontana et al. 1991; Fontana et al. 1993; Fontana and Schuster 1998; Fontana et al. 1993; Gruener et al. 1996a, 1996b) showed the following:

1. A large fraction of point mutations are neutral in RNA molecules in the sense that the mutation does not change the base pairing pattern (secondary structure) of the ground state structure.
2. The pre-image $\Phi^{-1}(y)$, i.e., the sequences folding into a common RNA structure y, is to a first approximation homogeneously distributed among the sequences C (y) that satisfy the base pairing constraints imposed by y. Note that $\Phi^{-1}(y) \subseteq C(y)$.
3. As a consequence of the high degree of neutrality (1) and the approximate homogeneity (2), there are extensive so-called "*neutral networks*" of sequences folding into the same ground state structure. These neutral networks "percolate" through sequence space and contain neutral paths that connect sequences without detectable sequence similarity.
4. The neutral networks tend to be connected or at least to decompose into only a small number of very large components.
5. The intersection theorem (Reidys et al. 1997) guarantees that the sets $C(y')$ and $C(y'')$ of sequences that are compatible with two arbitrary structures y' and y'' have non-empty intersection.
6. The neutral networks $\Phi^{-1}(y')$ and $\Phi^{-1}(y'')$ therefore come very close to each other, and the distance of an arbitrary sequence x_0 to a sequence $x \in \Phi^{-1}(y)$ folding into y is determined essentially by the violations of the base pairing constraints in x_0 only. This property is known as *shape space covering*.

This rather special structure of the RNA folding map implies a diffusion-like behavior of evolving populations of RNA molecules in sequence space, which conforms to Kimura's neutral theory (Huynen et al. 1996; Kimura 1983). It also implies constant rates of encountering novel variants along evolutionary trajectories (Huynen 1996). Thus, it explains the punctuated-equilibrium-like dynamics of RNA evolution characterized by long phases of diffusion on neutral networks interrupted by intermittent bursts of adaptive evolution when fitter mutants are encountered at the fringes of the neutral network (Huynen et al. 1996).

A beautiful illustration of these properties of the RNA folding map is the construction of a bistable ribozyme (Schultes and Bartel 2000): A single RNA folds into either of two evolutionarily unrelated ribozyme structures and catalyzes the corresponding reactions. Nevertheless, the bistable sequence has neighbors that are efficient catalysts for only one of the two alternative reactions and that are connected by neutral paths of the corresponding wild-type ribozyme.

Recent empirical work on very small RNA fitness landscapes defined by aptamer binding affinities, on the other hand, seems to be at odds with these observations and rather indicates as rugged structure without extensive paths (Athavale et al. 2014). The empirical aptamer binding landscape, however, focussed on the small subsequence actually involved in binding. This is indeed expected to be dominated by a few peaks corresponding to the best-binding motives. It forms a low-dimensional subspace that constrains the RNA molecules to those that have the binding motive but does not speak to the landscape defined by the large rest of the molecule. Computational studies indeed have shown that small sequence motives (as models of active sites or binding pockets) can be constrained without affecting the overall structure of the RNA folding map.

A different line of evidence for (nearly) neutral paths in RNA fitness landscapes comes from the comparative analysis homologous RNAs. Here we often observe a strong conservation of secondary structure while the sequence may have diverged already to beyond the detection limit (Torarinsson et al. 2006). Computational surveys with different methods have provided good evidence that this is not at all a rare phenomenon: More than 10 % of mammalian genomes are under stabilizing selection for RNA secondary structure elements, but more than 85 % of these elements are essentially unconstrained at sequence level (Smith et al. 2013). This provides a rather direct way to observe diffusive evolution on neutral networks.

2 Autocatalytic Networks

2.1 The Bioinformatics of RNA–RNA Interactions

Specific interactions among distinct RNA molecules are readily established by complementary base pairing, i.e., using the same principles that lead to the formation of intramolecular secondary structures. Conceptually straightforward (but computationally at times difficult) extensions of the secondary structure model thus also incorporate RNA–RNA interactions. Since there is no local difference between intermolecular and intramolecular base pairs, even the same energy parameters can be used. For an in-depth discussion, we refer to Backofen (2014) for a recent review. An important issue is the concentration dependence. In thermodynamic equilibrium, setting is expressed as

$$\frac{[AB]}{[A][B]} = K_{AB} = \frac{Z'_{AB}}{Z_A Z_B} \tag{11}$$

where $[A]$, $[B]$, and $[AB]$ are the concentrations of monomers of two types of RNAs and their dimer, respectively. The key observation is that the equilibrium constant K_{AB} can be computed directly from the sequences using the partition function versions of `RNAfold` (Z_A, Z_B) and `RNAcofold` or `RIP` (Z_{AB}). The correct partition function for the duplex can be expressed as $Z'_{AB} = (Z_{AB} - Z_A Z_B)e^{-\beta e}$, where Z_{AB} is the partition function computed directly by a cofolding approach, which also contains non-interacting conformations, and e is an initialization energy parameter capturing the additional entropic effects of forming a duplex (Bernhart et al. 2006; Dimitrov and Zuker 2004).

RNA–RNA interactions by means of canonical base pairing play an important role in post-transcriptional regulation. For example, the interaction of microRNAs with messenger RNAs and of small nucleolar RNAs with ribosomal RNAs is of this type. Similarly, it is the preferred mode of action of bacterial small RNAs. Of course, product inhibition in model systems of replicating nucleic acids is also owed to RNA–RNA binding *in trans*.

2.2 Replicator Networks

Eigen and Schuster noticed already in the late 1970s that systems of replicating molecules behave qualitatively different depending on whether the catalyst **E** in Eq. (7) is considered part of the environment or whether the replicators also catalyze their—or each other's—replication (Eigen and Schuster 1979). The net reaction of such a system can be abstracted in the form

$$\mathbf{A} + 2\mathbf{Y} \xrightarrow{k} 3\mathbf{Y} \tag{12}$$

for a single self-replicator with perfect accuracy and

$$\mathbf{A} + \mathbf{Y} + \mathbf{Z} \xrightarrow{q_{xy}k_{yz}} \mathbf{Y} + \mathbf{Z} + \mathbf{X} \tag{13}$$

in the general case of cross-catalysis with imperfect copying. As in the quasispecies model, **Y** is the template, **X**, which may coincide with **Y,** is the product of the copy reaction, and **Z** is the ribozyme catalyst. The kinetic constants k_{yz} describe the rate of copying template **Y** by catalyst **Z**. Again q_{xy} is the mutation probability of producing the offspring **X** from the template **Y**.

The construction of an RNA replicase ribozyme that is capable of copying a broad range of templates, including itself, has been an open problem for decades, ever since the discovery that RNA molecules have catalytic activities akin to proteins. As proof of principle, an RNA ligase ribozyme (Paul and Joyce 2002) was obtained in 2002, the first RNA replicase followed in 2009 (Lincoln and Joyce 2009), and was improved stepwise (Ferretti and Joyce 2013). Earlier this year, Roberson and Joyce finally described a self-replicating ribozyme that can sustain exponential growth (Robertson and Joyce 2014). It also copies a partner ribozyme so that the coupled system is capable of Darwinian evolution. Autocatalytic self-replicators of this type are in principle capable of open-ended Darwinian evolution. Experimental exploration of the test tube models of a hypothetical RNA world comprising autonomous interacting self-replicating RNAs thus is becoming feasible.

The dynamics of such a system can again be derived from the reaction mechanism under the assumption of mass action kinetics. In the simplest instantiation, it is of the form (Stadler and Schuster 1992)

$$\frac{\mathrm{d}[x]}{\mathrm{d}t} = \underbrace{[x]\left(\sum_{z} k_{xz}[z] - \varphi\right)}_{\text{selection}} + \underbrace{\sum_{y,z}\left(q_{xy}k_{y}z[y][z] - q_{yx}k_{xz}[y][z]\right)}_{\text{mutation}} \qquad (14)$$

Constant organization is enforced by balancing the flux with the net production in the system, i.e., $\varphi = \sum_{xz} k_{xz}[x][z]$, independent of the choice of the mutation rates q_{xy}. In the absence of mutation, $p \to 0$, the second term, which is proportional to the mutation rate p, vanishes. The remaining dynamical systems are known as the (quadratic) *replicator equation* (Schuster and Sigmund 1983). Maybe the most famous special case is the hypercycle model (Eigen and Schuster 1979). Second-order replicator equation also serves as a canonical model in game dynamics, and they are equivalent to the Lotka–Volterra equations, one of the first models of predator–prey interactions. They admit a rich mathematical theory with warrants books entirely dedicated to their analysis (Hofbauer and Sigmund 1998). More realistic reaction mechanisms again include variable levels of product inhibition. As in the quasispecies case, they are dominated by product inhibition for large total concentrations and eventually lead to global coexistence, i.e., the absence of selection. For moderate concentrations, however, complex dynamics described effectively by the catalyzed replication prevails Stadler et al. (2000).

2.3 Evolution of Autocatalytic Networks

Comparably, little is known about the evolution of sequences of autocatalytic networks. In Stadler (2002), the diffusion (in sequence space) of a population of interacting replicators has been studied. A major issue in models of this type is the assignment of the reaction rates k_{xy} as a function of sequences of template and catalyst. This problem is analogous to the assignment of fitness values to individual sequences in the quasispecies equation. It is even more challenging, however, since (1) we now have a quadratic number of coefficients to determine and (2) virtually nothing is known about the sequence dependence of the catalytic capabilities of nucleic acids. Very simple minimal models thus have been used: In Stadler (2002), k_{xy} is assumed to be dependent on the Hamming distance of x and y. In Forst (2000) and later in Stephan-Otto Attolini and Stadler (2006), the reaction rates were assumed to depend on the interaction structure of x and z as computed by `RNAcofold`. Neutrality in the interaction structure, i.e., of `RNAcofold` (x, z) w.r.t. mutation in either x or y, is important for evolvability in sequence space as well as the persistence of the population (Stephan-Otto Attolini and Stadler 2006). It remains unknown, however, whether this type of model is realistic even in a statistical sense. One would assume that the catalytic activity of a catalyst z on a template z depends on local interactions close to the processive site rather than on a conserved global structure.

2.4 Distributed Autocatalysis

Macromolecules that are directly self-replicating, i.e., that can copy a template including a second copy of themselves, are certainly the conceptually simplest building blocks of a self-propagating system. Since it has remained open for a long time whether RNA replicase enzymes can be constructed, alternative architectures have been explored at least theoretically. Assuming that copy machines are infeasible, systems of chemical reactions have been studied in which some of the chemical species also act as catalysts. Seminal work in this direction includes Stuart Kauffman's string concatenation model (Kauffman 1986) or Walter Fontana's artificial chemistry based on the lambda calculus (Fontana and Buss 1994). The key question is then to characterize closed, self-maintaining sets that collectively behave as an autocatalyst. Despite substantial progress in the mathematical and computational analysis of this class of models (Hordijk et al. 2012, 2014; Smith et al. 2014), it remains unclear whether and how they may have played a role in the origin of life. For example, it can be shown that in large chemical systems with n distinct molecular species, each molecule must catalyze $\propto \log n$ reaction in order to make it likely to find a collective autocatalytic set (Hordijk et al. 2011). At present, no plausible material instantiation appears to be known, and it remains to be seen whether the required abundance and specificity of catalytic activities are realistic for some kind of chemistry.

References

Acevedo A, Brodsky L, Andino R (2014) Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature 505:686–690

Aita T, Hamamatsu N, Nomiya Y, Uchiyama H, Shibanaka Y, Husimi Y (2002) Surveying a local fitness landscape of a protein with epistatic sites for the study of directed evolution. Biopolymers 64:95–105

Aita T, Uchiyama H, Inaoka T, Nakajima M, Kokubo T, Husimi Y (2000) Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type landscape: application to prolyl endopeptidase and thermolysin. Biopolymers 54:64–79

Athavale SS, Spicer B, Chen IA (2014) Experimental fitness landscapes to understand the molecular evolution of RNA-based life. Curr Opin Chem Biol 22C:35–39

Babajide A, Hofacker IL, Sippl MJ, Stadler PF (1997) Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. Fold Des 2:261–269

Backofen R (2014) Computational prediction of RNA–RNA interactions. Methods Mol Biol 1097:417–435

Bernhart SH, Tafer H, Mückstein U, Flamm C, Stadler PF, Hofacker IL (2006) Partition function and base pairing probabilities of RNA heterodimers. Algorithms Mol Biol 1:3 (epub)

Biebricher CK, Eigen M (1988) Kinetics of RNA replication by *Qβ* replicase. In: Domingo E, Holland JJ, Ahlquist P (eds) RNA genetics. RNA directed virus replication, vol I. CRC Press, Boca Raton, FL, pp 1–21

Biebricher CK, Luce R (1992) In vitro recombination and terminal elongation of RNA by *Qβ* replicase. EMBO J 11:5129–5135

Chan HS, Bornberg-Bauer E (2002) Perspectives on protein evolution from simple exact models. Appl Bioinf 1:121–144

Dimitrov RA, Zuker M (2004) Prediction of hybridization and melting for double-stranded nucleic acids. Biophys J 87:215–226

Eigen M (1971) Selforganization of matter and the evolution of macromolecules. Naturwissenschaften 58:465–523

Eigen M, McCaskill J, Schuster P (1989) The molecular quasispecies. Adv Chem Phys 75:149–263

Eigen M, Schuster P (1977) The hypercycle. A principle of natural self-organization. Part A: emergence of the hypercycle. Naturwissenschaften 64:541–565

Eigen M, Schuster P (1979) The hypercycle. Springer, New York

Ellington AD, Szostak JW (1990) In vitro selection of RNA molecules that bind specific ligands. Nature 346:818–822

Erlich HA (ed) (1989) PCR technology. Principles and applications for DNA amplification. Stockton Press, New York

Fahy E, Kwoh DY, Gingeras TR (1991) Self-sustained sequence replication (3SR): an isothermal transcription-based amplification system alternative to PCR. PCR Methods Appl 1:25–33

Ferretti AC, Joyce GF (2013) Kinetic properties of an RNA enzyme that undergoes self-sustained exponential amplification. Biochemistry 52:1227–1235

Flamm C, Stadler BMR, Stadler PF (2007) Saddles and barrier in landscapes of generalized search operators. In: Stephens CR, Toussaint M, Whitley D, Stadler PF (eds) Foundations of Genetic Algortithms IX. Lecture Notes Computer Science. 9th International Workshop, FOGA 2007, Mexico City, Mexico, vol 4436. Springer, Berlin, Heidelberg, 8–11 Jan 2007, pp 194–212

Fontana W, Buss LW (1994) What would be conserved "if the tape were played twice". Proc Natl Acad Sci USA 91:757–761

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

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

Fontana W, Schuster P (1998) Continuity in evolution: on the nature of transitions. Science 280:1451–1455

Fontana W, Stadler PF, Bornberg-Bauer EG, Griesmacher T, Hofacker IL, Tacker M, Tarazona P, Weinberger ED, Schuster P (1993b) RNA folding landscapes and combinatory landscapes. Phys Rev E 47:2083–2099

Forst CV (2000) Molecular evolution of catalysis. J Theor Biol 205:409–431

Gilbert W (1986) The RNA world. Nature 319:618

Gorodkin J, Ruzzo WL (2014) RNA sequence, structure, and function: computational and bioinformatic methods. Humana Press, New York City

Gruener W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Stadler PF, Schuster P (1996a) Analysis of RNA sequence structure maps by exhaustive enumeration. I neutral networks. Monath Chem 127:355–374

Gruener W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Stadler PF, Schuster P (1996b) Analysis of RNA sequence structure maps by exhaustive enumeration. II. structures of neutral networks and shape space covering. Monath Chem 127:375–389

Happel R, Stadler PF (1999) Autocatalytic replication in a CSTR and constant organization. J Math Biol 38:422–434

Hayashi Y, Aita T, Toyota H, Husimi Y, Urabe I, Yomo T (2006) Experimental rugged fitness landscape in protein sequence space. PLoS ONE 1:e96

Hietpas RT, Jensen JD, Bolon DN (2011) Experimental illumination of a fitness landscape. Proc Natl Acad Sci USA 108:7896–7901

Hofbauer J, Sigmund K (1998) Dynamical systems and the theory of evolution. Cambridge University Press, Cambridge

Hordijk W, Kauffman SA, Steel M (2011) Required levels of catalysis for emergence of autocatalytic sets in models of chemical reaction systems. Int J Mol Sci 12:3085–3101

Hordijk W, Steel M, Kauffman S (2012) The structure of autocatalytic sets: evolvability, enablement, and emergence. Acta Biotheor 60:379–392

Hordijk W, Wills PR, Steel MA (2014) Autocatalytic sets and biological specificity. Bull Math Biol 76:201–224

Huynen MA (1996) Exploring phenotype space through neutral evolution. J Mol Evol 43:165–169

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

Jimenez JI, Xulvi-Brunet R, Campbell G, Turk-MacLeod R, Chen IA (2013) Comprehensive experimental fitness landscape and evolutionary network for small RNA. Proc Natl Acad Sci USA 110:14984–14989

Kauffman S (1986) Autocatalytic sets of proteins. J Theor Biol 119:1–24

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kouyos RD, Leventhal GE, Hinkley T, Haddad M, Whitcomb JM, Petropoulos CJ, Bonhoeffer S (2012) Exploring the complexity of the HIV-1 fitness landscape. PLoS Genet 8:e1002551

Kramer FR, Mills DR, Cole PE, Nishihara T, Spiegelman S (1974) Evolution in vitro: sequence and phenotype of a mutant RNA resistant to ethidium bromide. J Mol Biol 89:719–736

Lauring AS, Andino R (2011) Exploring the fitness landscape of an RNA virus by using a universal barcode microarray. J Virol 85:3780–3791

Lee DH, Granja JR, Martinez JA, Severin K, Ghadiri MR (1996) A self-replicating peptide. Nature 382:525–528

Levine HA, Nilsen-Hamilton M (2007) A mathematical analysis of SELEX. Comp Biol Chem 31:11–35

Li T, Nicolaou KC (1994) Chemical self-replication of palindromic duplex DNA. Nature 369:218–221

Lincoln TA, Joyce GF (2009) Self-sustained replication of an RNA enzyme. Science 323:1229–1232

Lobkovsky AE, Wolf Y, Koonin EV (2011) Predictability of evolutionary trajectories in fitness landscapes. PLoS Comput Biol 7:e1002302

Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL (2011) ViennaRNA package 2.0. Alg Mol Biol 6:26

Luthra R, Medeiros LJ (2004) Isothermal multiple displacement amplification: a highly reliable approach for generating unlimited high molecular weight genomic DNA from clinical specimens. J Mol Diagn 6:236–242

Mullis KB, Faloona F, Scharf S, Saiki R, Horn G, Erlich H (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 51(1):263–273

Otwinowski J, Plotkin JB (2014) Inferring fitness landscapes by regression produces biased estimates of epistasis. Proc Natl Acad Sci USA 111:E2301–E2309

Paul N, Joyce GF (2002) A self-replicating ligase ribozyme. Proc Natl Acad Sci USA 99:12733–12740

Pitt JN, Ferré-D'Amaré AR (2010) Rapid construction of empirical RNA fitness landscapes. Science 330:376–379

Plöger TA, Kiedrowski G (2014) A self-replicating peptide nucleic acid. Org Biomol Chem 12:6908–6914

Rasmussen S, Chen L, Nilsson M, Abe S (2003) Bridging nonliving to living matter. Artif Life 9:269–316

Reetz MT, Sanchis J (2008) Constructing and analyzing the fitness landscape of an experimental evolutionary process. ChemBioChem 9:2260–2267

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

Robertson MP, Joyce GF (2014) Highly efficient self-replicating RNA enzymes. Chem Biol 21:238–245

Romero PA, Krause A, Arnold FH (2013) Navigating the protein fitness landscape with Gaussian processes. Proc Natl Acad Sci USA 110:E193–E201

Rowe W, Platt M, Wedge DC, Day PJ, Kell DB, Knowles J (2010) Analysis of a complete DNA-protein affinity landscape. J R Soc Interface 7:397–408

Schultes EA, Bartel DP (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds. Science 289:448–452

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

Schuster P, Sigmund K (1983) Replicator dynamics. J Theor Biol 100:533–538

Segel LA, Slemrod M (1989) The quasi-steady state assumption: a case study in perturbation. SIAM Rev 31:446–477

Smith JI, Steel M, Hordijk W (2014) Autocatalytic sets in a partitioned biochemical network. J Syst Chem 5:2

Smith MA, Gesell T, Stadler PF, Mattick JS (2013) Widespread purifying selection on RNA structure in mammals. Nucleic Acids Res 41:8220–8236

Stadler BMR (2002) Diffusion of a population of interacting replicators in sequence space. Adv Complex Syst 5(4):457–461

Stadler BMR, Stadler PF, Schuster P (2000) Dynamics of autocatalytic replicator networks based on higher order ligation reactions. Bull Math Biol 62:1061–1086

Stadler PF, Schuster P (1992) Mutation in autocatalytic networks—an analysis based on perturbation theory. J Math Biol 30:597–631

Stephan-Otto Attolini C, Stadler PF (2006) Evolving towards the hypercycle: a spatial model of molecular evolution. Physica D 217:134–141

Szathmáry E, Gladkih I (1989) Sub-exponential growth and coexistence of non-enzymatically replicating templates. J Theor Biol 138:55–58

Torarinsson E, Sawera M, Havgaard JH, Fredholm M, Gorodkin J (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. Genome Res 16:885–889

Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249:505–510

Turner DH, Mathews DH (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res 38D:280–282

Varga S, Szathmáry E (1997) An extremum principle for parabolic competition. Bull Math Biol 59:1145–1154

von Kiedrowski G (1986) A self-replicating hexadeoxynucleotide. Angew Chem Int Ed Engl 25:932–935

Wills PR, Kauffman SA, Stadler BM, Stadler PF (1998) Selection dynamics in autocatalytic systems: templates replicating through binary ligation. Bull Math Biol 60:1073–1098

Woo HJ, Reifman J (2014) Quantitative modeling of virus evolutionary dynamics and adaptation in serial passages using empirically inferred fitness landscapes. J Virol 88:1039–1050

Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9:133–148