Success and efficiency of Darwinian evolution is based on the dichotomy of genotype and phenotype: The former is the object under variation whereas the latter constitutes the target of selection. Genotype-phenotype relations are highly complex and hence variation and selection appear uncorrelated. Population genetics visualizes evolutionary dynamics as a process among genotypes. Phenotypes are represented through empirical parameters only. The quasispecies concept introduces the molecular mechanism of mutation. Optimization is seen as a process in genotype space. Populations optimize through adaptive walks. Selective neutrality leads to random drift. Understanding evolution will be always incomplete unless phenotypes are considered explicitly. At the current state of the art, almost all genotype-phenotype mappings are too complex to be analyzed and modeled. Only the most simple case of an evolutionary process, the optimization of RNA molecules \textit{in vitro}, where genotypes and phenotypes are RNA sequences and structures, respectively, can be treated successfully. We derive a model based on a stochastic process which includes unfolding of genotypes to form phenotypes as well as their evaluation. Relations between genotypes and phenotypes are handled as mappings from sequence space onto the space of molecular structures. Generic properties of this map are analyzed for RNA secondary structures. Optimization of molecular properties in populations is modeled \textit{in silico} through replication and mutation in a flow reactor. The approach towards a predefined structure is monitored and reconstructed in terms of an uninterrupted series of phenotypes from initial structure to target, called relay series. We give a novel definition of continuity in evolution which identifies discontinuities as major changes in molecular phenotypes.
1 GENOTYPES AND PHENOTYPES

Evolutionary optimization in asexually multiplying populations follows Darwin’s principle and is determined by the interplay of two processes which exert counteracting influences on genetic heterogeneity: (i) Mutations increase diversity of genotypes, and (ii) selection decreases diversity of phenotypes. Recombination occurring obligatorily in sexually reproducing populations is another process contributing to the maintenance of diversity. The genotypes of the offspring combine parts from both parental genotypes. Variation and selection operate on different manifestations of the individual, genotype and phenotype, respectively. At the first glance, decoupling of the targets of mutation (or recombination) and selection may seem to be a disadvantage. As a result of uncorrelatedness an advantageous mutation does not occur more frequently because it has a better chance to become selected. Considering non-biological complex optimization problems, however, random variation is well known to be a powerful strategy [65]. Deterministic optimization techniques, gradient techniques, for example, are too easily caught in local extrema and can neither approach optimal nor near optimal solutions on multipeak landscapes derived from sophisticated cost functions. Genotype-phenotype dichotomy in nature guarantees randomness of moves in Darwinian optimization.

Separation of genotype and phenotype is trivially fulfilled in higher forms of life where the phenotype is an adult multicellular organism created through development which unfolds the genotype in a manner that reminds of the execution of a computer program. Genotype and phenotype are different entities and with the exception of few examples it is currently impossible to infer changes in phenotypic properties from known modifications in the DNA sequence of the genotype. In addition, epigenetics exerts influence on the phenotype which are by definition distinct from genetics. In case of unicellular organisms, procaryotic or eucaryotic, the phenotype comprises cellular metabolism in its full complexity. In today’s reality, metabolism is too complex to be deduced from the genomic DNA sequence. Again genotype and phenotype are clearly distinct features of the organism. In vitro evolution deals also with Darwinian optimization in populations of molecules which are capable of replication. Here, the distinction between genotype and phenotype is more subtle and therefore we shall consider these experiments in more detail.

Sol Spiegelman and his group [92] pioneered experiments on evolution of RNA molecules in the test-tube (figure 1). A sample of RNA of the bacteriophage Qβ was transferred into a solution containing an RNA replicase and the
FIGURE 1  **Stepwise increase in the rate of RNA production.** The upper part shows the technique of serial transfer applied to evolution of RNA molecules in the test tube. The material consumed in the synthesis is replaced by transfer of a small sample into a new test tube with fresh stock solution. The stock solution contains an enzyme required for replication, $Q\beta$-replicase, for example, and the activated monomers (ATP, UTP, GTP, and CTP), the building blocks for polynucleotide synthesis. The rate of RNA production (lower part) is measured through incorporation of radioactive GTP into the newly synthesized RNA molecules. The figure is redrawn from the data in [66].
activated monomers for RNA synthesis (ATP, UTP, GTP, and CTP). RNA replication sets in and the material in the solution is consumed. After some time the consumed material is replaced through transfer of a small sample into fresh stock solution. This procedure is repeated some fifty to hundred times. In such serial transfer experiments the rate of RNA synthesis increases by more than one order of magnitude. Spiegelman identified the nucleotide sequence of an RNA molecule as its genotype and the molecular structure as its phenotype. Genotype and phenotype thus are two different manifestations of the same molecule, known to the biochemist as sequence and spatial structure, respectively. Here, we cannot be sure \textit{a priori} that genotype and phenotype are sufficiently distinct in order to lead to \textit{de facto} uncorrelatedness of mutation and selection. Considering RNA folding in detail, however, we realize that structure formation is a highly complex process that does not (yet) generally allow to infer structural changes from mutations in the sequence. At best we have to go through a sophisticated algorithm that predicts structure from known sequence (figure 2). A characteristic of sequence-structure relations is that small changes in sequence may but need not have small consequences for structure, and at a (sufficiently) coarse grained level, the sequence-structure map appears to be almost random (see Sect. 3). All three examples discussed above do not allow for a direct feedback mechanism which translates possible consequences of mutation into the frequency of mutant formation. In the absence of such a feedback random choice of moves is certainly the best strategy.

Different notions of structure imply different models for the molecular phenotype. Examples are: (i) the conformation of minimal free energy (mfe) which is formed after sufficiently long enough time and at sufficiently low temperature, (ii) the mfe structure together with Boltzmann weighted suboptimal conformations in the sense of a partition function, and (iii) kinetic structures or ensembles of structures which take available folding times into account and acknowledge the fact that RNA is produced in the cell through transcription that forms the newly synthesized RNA strand from 5'-end to 3'-end. In thermodynamic language we call the mfe conformation the structure in the limit $T \to 0K$ and $t \to \infty$ (i). The partition function represents the ensemble of structures at finite temperature $T$ and $t \to \infty$ (ii). A kinetic structure finally is defined for finite temperature and finite time (iii). It is commonly assumed that small RNA molecules form mfe structures on folding, but recent studies using of a new algorithm which resolves structure formation to formation and cleavage of single base pairs have shown that this is not necessarily true: Kinetic structures in the sense of metastable suboptimal conformations may play a role also for rather small RNA molecules [28]. For longer RNA sequences the discrepancy between most stable and kinetically favored metastable structures is well established [71]. Refolding kinetics of RNA structures shows that only sufficiently low barriers between mfe structures and metastable conformations can be passed at room temperature. If high barriers separate different valleys of the conformational landscape we observe only the subset of confor-
FIGURE 2 Folding of RNA sequences into structures. The folding is performed in two steps from the sequence to the secondary structure and from the secondary structure to the full spatial structure. The example shown is the transfer RNA molecule tRNA\textsubscript{phe}. Both steps occur under the condition of minimal free energy (mfe). The secondary structure is commonly defined as a listing of base pairs which is compatible with a planar graph without knots or pseudoknots.
infrequent incident. We shall try to give answers which are compatible with the now well established neutral [55] or nearly neutral [75] theory of evolution. The mean generation time is the time unit of evolution. It decides whether or not evolution experiments are feasible and can be carried through in reasonable times. Higher organisms have generation times from several weeks to more than a decade. The time spans required for direct observation of evolutionary phenomena are at least hundreds to thousands of years and thus too long for experiments. At present we are thus confined with three experimental systems to study evolution: polynucleotide molecules, viroids or viruses, and bacteria.

In order to set the stage for the development of a comprehensive theory of evolution we describe a particularly illustrative experiment. Bacteria are well suited objects for studies on evolution because generation times can be as short as 20 minutes under optimal conditions. The rate of mutation was determined for many DNA based microbes and, interestingly, it was found to have a constant value of about 0.0033 per genome and generation independently of DNA chain length [17]. An elegant serial transfer experiment with *Escherichia coli* bacteria was carried out by Richard Lenski and coworkers [26,57,76]. Populations of $5 \times 10^8$ cells were diluted 1:100 every day and recorded for about three years leading to about 10,000 generations which is tantamount to an average generation time of 3.6 hours. Fitness measured in terms of growth rate increased by about 40% during a fast adaptive period over the first 2,000 generations. This increase occurs in steps and not continuously as one might have expected [26]. After an adaptive initial period the curve saturates in the remaining 8,000 generations and settles on a plateau at about 1.5 times the initial fitness. More recently the rate of phenotypic evolution as monitored via fitness or cell size was complemented and compared with the rate of genomic evolution determined through DNA fingerprinting [76]: Phenotypic evolution is fast in the initial phase and slows down during saturation. Evolution of the genome, on the other hand, behaves differently: it does not slow down and seems even to speed up in the saturation phase. Although the values from experiments on two *E. coli* variants differ substantially, it is certain that the rate of genotypic change does not decrease during saturation as phenotypic evolution does. This finding asks indeed for an explanation since a comprehensive theory of evolution should be in a position to make correct predictions on relative rates of genomic and phenotypic evolution.

## 2 A WORLD OF GENOTYPES

The theory of population genetics was conceived and built by the three famous scholars, Ronald Fisher, John Haldane, and Sewall Wright, and united...
FIGURE 3  Evolution of genotype frequencies in asexual populations. The sketch shows typical solutions curves representing relative concentrations or frequencies \(x_i(t)\) of a population modeled by equation (1). New variants are formed by rare mutation events. Depending on replication rates relative to the mean value the frequencies will increase \((a_i > \bar{a})\), decrease \((a_i < \bar{a})\) or drift randomly in the neutral case \((a_i \approx \bar{a})\). Stochastic theory shows that fixation of mutants occurs also in neutral evolution: According to Kimura’s theory [55] the mean time from the appearance of a mutant, \(x_m(0) = 1/N\), until its fixation in the population, \(x_m(\tau_F) \approx 1\), is \(< \tau_F > = 2N\).

the previously conflicting issues of Darwinian evolution and Mendelian genetics in an elegant and straightforward way. Evolution is considered as an optimization process at the level of populations and the relevant variables are the frequencies of genes. The properties of phenotypes enter the model equations as parameters. Such parameters are, among others, life times, litter sizes, survival probabilities of descendants, and rates of reproduction, all of them contributing to fitness values. The notion of “optimization process” implies the definition of a direction: Every change or “move” along the defined direction is accepted, every move in opposite direction is rejected.\(^3\) The reduction of the phenotype to a set of input parameters for the differential equations of population dynamics is the basis of success and, at the same time, the most serious limitation of conventional population genetics. Proper choice of parameter values allows to model and analyze typical idealized situations and to study the influences of quantities like, for example, relative fitness, mutation rate, recombination rate or population size on the spreading of genes in popu-

\(^3\)Acception and rejection may be bound by predefined probability limits. Nothing is said so far about moves which are neither associated with progress nor with regression. Such “neutral” moves will be the subject of forthcoming sections.
Problem evolution. Problems arise when it becomes necessary to assign realistic values to the parameters, which are commonly very hard to determine experimentally, or when one aims at studies that deal with phenotypes explicitly. In the latter case we require knowledge on the relation between genotypes and phenotypes in order to be able to derive or model the consequences of changes in the genomic nucleotide sequence for the phenotype. Genotype-phenotyp maps are highly complex and hard to investigate even in the most simple cases. We shall discuss the particularly simple example of phenotypes in test-tube evolution being represented by the spatial structures of RNA molecules in the next Sect. 3.

2.1 Selection Equation

It is straightforward to model selection in populations with asexual replication. Since there is little or no recombination, the appropriate variables are numbers ($N_k$), concentrations ([I_k]), or frequencies ($x_k$) of genotypes $I_k$ rather than genes. The definitions are: $N_k = \#(I_k)$, the population size $N = \sum_{j=1}^{n} N_j$, $[I_k] = \#(I_k) / (V \cdot N_L)$ with $V$ being the reaction volume and $N_L$ Avogadro’s number, and $x_k = [I_k] / \sum_{j=1}^{n} [I_j]$. Suitable conditions for selection are provided, for example, by serial transfer experiments or by a flow reactor as discussed later on (figure 6). A constraint known as constant organization [24] is closely related to that of a flow reactor and leads to constant population size. Then, the normalization condition for the frequencies of genotypes, $\sum_{j=1}^{n} x_j = 1$, is readily incorporated into the differential equation describing the time dependence of genotype distributions in the limit of infinite population size:

$$\frac{dx_k}{dt} = x_k \left( e_k - \Phi(t) \right), \quad k = 1, \ldots, n.$$  

Herein $e_k = a_k - d_k$ is the net production rate constant of genotype $I_k$, which is obtained as the difference between replication rate constant $a_k$ and degradation rate constant $d_k$, and $\Phi(t) = \bar{e}(t) = \sum_{j=1}^{n} e_j x_j(t)$ is the mean net production which is tantamount to the excess reproduction rate of the population. Accordingly, we have $\sum_{j=1}^{n} dx_j/dt = 0$ leading to constant population size. Population geneticists measure progeny in terms of fitness. In the rare mutation case fitness is identical to net production: $f_k = e_k$ and $\Phi(t) = \bar{f}(t) = \sum_{j=1}^{n} f_j x_j(t)$. For equal degradation rates or life times, $\tau_k = \ln 2/d_k$, the contributions of degradation rates to $e_k$ and $\Phi(t)$ compensate each other exactly and the fitness values are equal to the rate constants of replication, $f_k = a_k$. Then, the input parameters of the equations of populations genetics (1), i.e. the parameters mentioned above, are simply the replication rate constants $a_k$ of the molecules, viruses or microorganisms. They are determined, in essence, by the corresponding phenotypes, molecular structures, viral life cycles or cellular metabolism, respectively.
Mutations are assumed to be rare events and they are not considered explicitly in the differential equations. At finite population size fluctuations become important. In addition, every mutant has to start from a single copy and hence it is jeopardized by random elimination. In order to account for random events selection in finite populations has been modeled by means of a master equation [20, 52, 58, 59, 95]. It turned out, however, that the convenient constraint of constant organization, as applied in equation (1), leads to instability in the sense that fluctuations in population size increase in time without limit. Two different modifications were applied which stabilize the stochastic selection equation:

(i) every random replication event is strongly combined to a random dilution event, which is tantamount to two-component elementary steps leaving the population size $N$ strictly constant [70], and
(ii) the assumption of a dilution flux $\Phi_0(t) = \sum_{j=1} a_j N_j / N_0$ with constant $N_0$. This dilution flux is consistent with a population size fluctuating around the fixed value $N_0$ in the stationary state [52].

The first case (i) is known as Moran model [70] and yields a strictly constant population size. The second approach (ii) corresponds to a populations size $N(t)$ with fluctuations are proportional to $\sqrt{N}$. In the limit of long times $N(t)$ approaches the constant $N_0$. Van Kampen’s expansion [101] was applied to derive stationary solutions of the stochastic selection problem [59]. Typical solution curves are shown in figure 3. Genotypes replace each other in the course of evolution. A typical snapshot will not show more than two genotypes at nonmarginal frequencies. Here we restrict ourselves to a brief presentation of results, which were derived from Motoo Kimura’s stochastic theory of neutral evolution [54, 55]. This notion was coined because Kimura’s concept allows to investigate the neutral case, $a_1 = \ldots = a_k = \bar{a}$.

In Kimura’s model the mean rate of evolution, $\langle k \rangle$, is measured as the number of mutant substitutions per generation time and can be expressed by

$$\langle k \rangle = N \cdot v \cdot u(N, s, p) = N \cdot v \cdot \frac{1 - \exp(-2Nsp)}{1 - \exp(-2Ns)} ,$$

where $N$ is again the population size, $v$ the mutation rate per genome and generation, and $u(N, s, p)$ represents the probability of fixation. This probability is a function of $s$, the selective advantage, and $p$, the initial frequency of the mutant. Since every mutant starts inevitably from a single copy we may put $p = 1/N$ and find

$$\langle k \rangle = N \cdot v \cdot \frac{1 - \exp(-2s)}{1 - \exp(-2Ns)} .$$

The selective advantage $s$ is measured relative to the neutral case: The fitness value is $f = f_0(1 + s)$ and thus neutrality implies $f = f_0$ or $s = 0$. 

---

4The selective advantage $s$ is measured relative to the neutral case: The fitness value is $f = f_0(1 + s)$ and thus neutrality implies $f = f_0$ or $s = 0$. 

In the neutral case the rate of evolution is computed to be \( < k > = v \), and the mean time of the replacement for a given genotype by the next one is \( < \tau_R > = 1/ < k > = v^{-1} \) generations. For substantial selective advantage, \( s > 1/(2N) \), we find \( < k > = 2Ns \cdot v \) since \( u \approx 2s \): The rate of evolution increases linearly with the selective advantage \( s \) and the population size \( N \). On the average, a genotype will be replaced by the next one after \( < \tau_R > = (2Ns \cdot v)^{-1} \) generations (which is always shorter than in the neutral case because \( s > 1/(2N) \) holds). At still higher values of the selective advantage \( s \) the probability of fixation converges to \( \lim_{s \to \infty} u(s) = 1 \) and thus we find in the limit of infinite advantage: \( \lim_{s \to \infty} < k > = N \cdot v \). Accordingly, the mean rate of evolution for neutral, weakly and strongly advantageous variants is confined by \( v \leq < k > \leq N \cdot v \). We can use this expressions for rough estimates on the time required for the observation of evolutionary phenomena.

It is also worth considering the mean time for fixation of neutral and advantageous variants, \( \tau_F \). The solution of the deterministic selection equation for two genotypes, \( x_2 = x \), \( x_1 = 1 - x \) and \( f_0 = 1 \), is readily obtained in analytical form:

\[
x(t) = \frac{x_0}{x_0 + (1 - x_0) \cdot \exp(-st)}.
\]

From this equation we compute \( \tau_F \) as the time it takes for a mutant to grow from a single copy, \( x(0) = 1 \), to population size, \( x(\tau_F) \geq N - 1 \) and find: \( \tau_F \approx 2 \ln N/s \). For sufficiently large selective advantage, \( s > 1/(2N) \), the mean time of fixation is substantially shorter than in the neutral case, where \( \tau_F = 2N \) (see figure 3). In other words, the inverse of the mean fixation time, \( \tau_F^{-1} \), decreases linearly with \( s \) in the deterministic limit \( s \to 0 \), but the term from neutral evolution guarantees that the reciprocal time of fixation does not fall below the limit \( \tau_F^{-1} = 1/2N \).

### 2.2 MOLECULAR QUASISPECIES

An extension of conventional population genetics which considers evolution as chemical reactions in genotype space was proposed by Manfred Eigen in his seminal paper on the theory of the evolution of molecules [21]. His concept can be understood, in essence, as an application of chemical reaction kinetics to molecular evolution. A main issue of Eigen’s approach was to derive the mechanism by which biological information is created. Populations migrate through sequence space as metastable but structured distributions of
FIGURE 4  A quasispecies-type mutant distribution around a master sequence. The quasispecies is an ordered distribution of polynucleotide sequences (RNA or DNA) in sequence space $T'_n$. A fittest genotype or master sequence $I_m$, which is present at highest concentration, is surrounded in sequence space by a "cloud" of closely related sequences. Relatedness of sequences is expressed (in terms of error classes) by the number of mutations which are required to produce them as mutants of the master sequence. In case of point mutations the distance between sequences is the Hamming distance.$^6$ In precise terms, the quasispecies is defined as the stable stationary solution of equation (3) $^{21, 24}$, the mutant distribution described by the largest eigenvector of the matrix $W = \{W_{ij} = Q_{ij} \cdot a_j; i, j = 1, \ldots, n\}$$^{51, 73, 96, 100}$ (Its diagonal elements are approximations for fitness values, $f_k: f_k \approx W_{kk} = a_k \cdot Q_{kk}$). In reality, such a stationary solution exist only if the error rate of replication lies below a maximal value called the error threshold. In this region, i.e. below the often sharply defined mutation rate of the error threshold, this eigenvector represents a structured population as shown in the figure. Above the critical error rate the largest eigenvector is (practically) identical with the uniform distribution. The uniform distribution, however, can never be realized in nature or in vitro since the number of possible nucleic acid sequences ($4^n$) exceeds the number of individuals by many orders of magnitude even in the largest populations. The actual behavior is determined by incorrect replication leading to random drift: populations migrate through sequence space.

genotypes and at the same time optimize mean fitness. Populations explore environments by means of a variation-selection process and gain information on them thereby. At the same time biological information is laid down in genotypes, being selected polynucleotide sequences of DNA or RNA. The de-
Evolution of Phenotypes

terministic equation (1) is readily extended to handle the frequent mutation scenario and we obtain the replication-mutation equation:

\[
\frac{dx_k}{dt} = x_k \left( Q_{kk} a_k - \Phi(t) \right) + \sum_{j=1,j\neq k}^{n} Q_{kj} a_j x_j, \quad k = 1, \ldots, n, \quad (3)
\]

with the mutation matrix \( Q = \{Q_{ij}; i, j = 1, \ldots, n\} \) and the mean excess production \( \Phi(t) = \bar{a}(t) = \sum_{j=1}^{n} a_j x_j(t) \) as before. In essence, the ansatz (3) differs from conventional population genetics as expressed, for example, by equation (1) in the handling of mutations. The conventional treatment introduces mutation as a rare stochastic event in the environment which is not controlled by the replication mechanism. Mutation is thus characterized by a probability density, commonly by an expectation value and, eventually, a variance. The replication-mutation equation (3), however, deals explicitly with mutations and handles error-free and incorrect replication as parallel reactions. Relative frequencies of the corresponding reaction channels are given by the elements of the mutation matrix. In particular, \( Q_{kj} \) is the frequency at which the genotype \( I_k \) is synthesized as an error-copy of \( I_j \). In general, equation (3) settles down to a stationary state corresponding to a stable or metastable\(^7\) distribution of genotypes. Such distributions of genotypes, called molecular quasispecies \([23, 24]\), are ordered: In the center of the distribution we find a most frequent and commonly also fittest genotype called the master sequence (figure 4). Quasispecies represent the genetic reservoirs of asexually replicating species, for example molecules, viroids, viruses, or bacteria.

Fitness landscapes are understood as distributions of fitness values over sequence space and can be represented by mappings from sequence space into the real numbers, \( g : \{I; d_{ij}^\ell\} \Rightarrow \mathbb{R}^\ell \) (For recent reviews see \([86, 93]\)). Herein the sequence space space is denoted by \( I \) and the distance between two sequences by \( d_{ij}^\ell \). Several classes of model landscapes were either invented, like for example the Nk-model \([53]\), or adopted from physical models of spin-glasses \([99]\) in order to study replication and mutation with distributions of fitness values in sequence space which are thought to be representative for real situations. Most landscapes sustain stable quasispecies only for mutation rates below a certain critical value called the error threshold \([23, 24]\). At the critical

\(6\) The symbol \( I_\ell \) is used here for genotypes or polynucleotide sequences (DNA, RNA) as well as \( I \) for the space of genotypes, called sequence space \([44]\) in order to point out that the sequences are the carriers of biological information. In particular, \( I_{\ell, \kappa}^\ell \) is the sequence space of sequences of chain length \( \ell \) over an alphabet of size \( \kappa \) (GC: \( \kappa = 2 \); AUGC: \( \kappa = 4 \)). \( I_{\ell, \kappa}^\ell \) carries \( \kappa^\ell \) different sequences. The Hamming distance \([43]\) of two sequences \( I_i \) and \( I_j \) is denoted by \( d_{ij}^\ell \) and counts the number of positions in which two aligned strings differ. It induces a metric on the corresponding sequence space.

\(7\) There are two reasons why the infinite time solution of equation (3) may become unstable: (i) The steady state can never be reached because the population size is too small to approach the infinite population limit, and (ii) random drift in the sense of neutral evolution (which is not addressed by deterministic equations) may be followed by a new onset of selection after an epoch of stasis, thereby causing the population to switch from one (quasispecies like) metastable state to another.
point an abrupt change in evolutionary dynamics is observed which reminds of a phase transition [60] (Exceptions are classes of artificially smooth landscapes sometimes applied in population genetics which show gradual transitions from quasispecies to uniform distributions). At error frequencies above threshold populations migrate through sequence space in random walk manner and do not approach stationary states (figure 4).

It is worth considering the parameter problem of population genetics once more, this time from the practical point of computer simulation. The numbers of possible RNA sequences or structures, the cardinalities of sequence space or shape space, are enormous, even for moderate chain length, and hence too large for any direct assignment of empirical parameters. Required are tractable models that allow to compute all relevant phenotypic quantities, in particular the $a_k$- and $Q_{kj}$-values, from rules which make use of a few parameters only.

As an example we consider a useful and realistic model for the mutation matrix $Q$ which is based on the notion of sequence space ($I_\ell$). Restriction to point mutations and assumption of uniform error rates, implying that the probability of a mutation is independent of the nature of the base exchange and the position along the sequence, allow to express all elements of $Q$ in terms of only three parameters only, $\ell$, $q$ and $d_{kj}$:

$$Q_{kj} = q^\ell \left(1 - q\right) d_{kj}.$$  (4)

Herein $\ell$ is the length of the polynucleotide chain. The single digit accuracy of replication $q$ implies a uniform error rate of $p = 1 - q$ per digit and replication event which is independent of the position in the sequence, and $d_{kj}$, finally, is the Hamming distance between the two sequences to be interconverted by the mutation.

Within the frame of the uniform error rate model (4) it is straightforward to compute an approximate expression of the critical mutation rate [21]. First, mutational backflow (the sum in the right hand part of equation 3) is neglected and second, the stationary frequency of the master genotype is computed to

$$\bar{x}_m = \left(\sigma_m Q_{mm} - 1\right) / \left(\sigma_m - 1\right),$$

wherein $\sigma_m = a_m / a_{k\neq m}$ defines the superiority of the master sequence and $\bar{a}_m = (\sum_{j=1,j\neq m}^n a_j x_j) / (1 - x_m)$ the mean replication rate of the population except the master. The expression for the error threshold is derived now by computing the error rate at which the concentration of the master sequence vanishes: $\bar{x}_m = 0 \rightarrow Q_{mm} = \sigma_m^{-1}$. Application of equation (4), $Q_{mm} = q^\ell$, yields two equations, one for the maximal mutation rate (or minimal replication accuracy, $p_{\text{max}} = 1 - q_{\text{min}}$) at constant chain length and one for the maximal chain length at constant mutation rate,

$$p_{\text{max}} = 1 - \sigma_m^{-1/\ell} \quad \text{and} \quad \ell_{\text{max}} = -\frac{\ln \sigma_m}{\ln q} \approx \frac{\ln \sigma_m}{1 - q} = \frac{\ln \sigma_m}{p},$$  (5)

which define the error threshold. Despite the simplifications made in the derivation of equation (5) the agreement between the exact curves $\bar{x}_m(q)$
and the approximation is surprisingly good [23]. For the sake of completeness we mention that the computation of an error threshold of replication and mutation has been extended to the diploid case [108] as well as to neutral evolution where stationarity refers to time independent distributions of phenotypes rather than genotypes [80]. Quasispecies formation has been studied also on a dynamical landscape [72].

Although all entries of the mutation matrix are now computable from a few input parameters, still more empirical data are required. As in the selection equation (1) the fitness values of phenotypes enter the kinetic differential equations (3) as parameters. Assignment of fitness values can be performed under model assumptions only. Considering, for example, a rather short RNA molecule of chain length $\ell = 100$ we are dealing with $1.6 \times 10^{60}$ different genotypes which may give rise to a smaller but still very large number of phenotypes. Commonly the problem is overcome by rather drastic simplifications. As an example we mention the single peak fitness landscape: One replication rate $a_m = \sigma$ is assigned to the fittest genotype, $I_m$, and all other genotypes are assumed to have replication rate $a_k \neq m = 1$ [96]. This ansatz reminds of the mean field approximation often used in physics. Since details of the distribution of fitness values are unknown, one replaces them by a mean value for all genotypes except the fittest one. In this sense, the single peak landscape has been used, for example, to derive analytical expressions for the threshold value of the error rate [21, 23, 24] (For further work on replication and mutation on model landscapes see [1, 2, 8, 9, 63, 72, 91, 99]).

An analysis of stochastic effects in replication-mutation systems based on multitype branching processes [12] provided a mathematical interpretation of the error threshold in terms first passage times: The probability of survival to infinite time of the master sequence is nonzero at error rates below threshold and becomes zero at the critical value. Above threshold, however, all genotypes have zero probability of survival to infinite time which is tantamount to instability of stationary sequence distributions. A later approach modeled replication and mutation as a birth-and-death process [73] and resulted in an analytical expression for the error threshold in finite populations:

$$q_{\min}(N) = q_{\min}(\infty) \left(1 + \frac{2\sqrt{\sigma_m - 1}}{\ell\sqrt{N}} + \ldots\right).$$

Other treatments of replication and mutation as stochastic processes were based on the corresponding master equation [20, 52, 58, 59, 95] and applied the same constraints as discussed for the selection equation, the Moran model [70] or the dilution flux $\Phi_0(t)$.

In summary, the quasispecies concept (3) provides a solution for the mutation problem but does not yet deal explicitly with phenotypes and their properties. The problem in handling phenotypes is twofold: First, the mapping of genotypes into phenotypes is extremely complicated and generally very hard to model. Second, phenotypes have a large number of properties most of which contribute to fitness only in an indirect way. What is needed therefore
is a realistic but sufficiently simple toy model that allows to compute fitness values from a set of few (simple) rules.

Stationary mutant distributions were observed and analyzed in the case of evolution in vitro of RNA molecules [5,83]. The data recorded in these studies reproduce well the predictions derived from equation (3). The threshold equation (5) can also be used to predict maximum genome lengths provided the population evolves at maximal mutation rate. This was indeed found to be the case with lytic RNA viruses [18] where the observed mutation rate $p$ increases linearly with the reciprocal chain length $\ell^{-1}$. A straightforward interpretation of this finding says that these viruses mutate as fast as possible because they have to cope with the powerful defense mechanisms of their hosts. A comparison of mutation rates in different groups of prokaryotes [17, 19] revealed a roughly constant mutation rate per genome length: $\ell \cdot p = \text{const}$. The constants are around 1 for lytic RNA viruses, roughly 0.1 for retroviruses and retrotransposons, and close to $1/300$ for microbes with DNA-based genomes (For an interpretation of these results on the basis of cost balance between reduction of deleterious mutants and precision of the replication machinery see [19]). In addition, the quasispecies concept turned out to be useful for the description of the evolution of RNA virus populations [14,15] and provided hints for the development of novel antiviral strategies [16].

2.3 EXTENSION TO PHENOTYPES

The first explicit consideration of phenotypes in a model of molecular evolution was implemented in silico in order to simulate replication and mutation in a flow reactor (figure 6) [31]. This simple model was already in a position to perform optimizations of RNA properties like thermodynamic stability or net productivity as expressed by the difference of replication and degradation rate constants ($e_k = a_k - d_k$). Later on, the relation between genotypes and phenotypes was made more precise and modeled as a mapping from sequence space into phenotype space. To this end we assume a metric phenotype space $S$ with some (hypothetical) measure of distance between phenotypes, $d_{ij}^S$:

$$\psi: \{I; d_{ij}^h\} \Rightarrow \{S; d_{ij}^S\}. \quad (6)$$

In other words, $S_k = \psi(I_k)$, implies that a phenotype $S_k$ is uniquely assigned to the genotype $I_k$. The assignment expressed by equation (6) is tantamount to the formation of the phenotype $S_k$ through unfolding of the genetic information stored in the genotype $I_k$. Fitness values are approximated by the product of replication rate constants and replication accuracy, $f_k \approx a_k \cdot Q_{kk}$, and can be seen as the result of a mapping $f$ from the phenotypes into the nonnegative real numbers:

$$f: \{S; d_{ij}^S\} \Rightarrow \mathbb{R}_+. \quad (7)$$

The map (7) evaluates the phenotype and returns its fitness value. In summary, we obtain fitness values from the genotype through the function:
\[ f(S_k) = f(\psi(I_k)) = f_k \approx a_k \cdot Q_{kk}. \]

The mapping \( \psi(\cdot) \), in general, cannot be expressed in analytical terms. At best we have algorithms that allow to compute structures from sequences (see next section 3). The situation is not less complex for the derivation of fitness values of phenotypes, but in this case the evaluation is often done by means of simple model functions. For example, \( f(\cdot) \) can be assumed to be a simple function of the distance between the phenotype and some target to be approached.

Now we are in a position to classify different mappings:

(i) \( \psi(\cdot) \) maps a discrete vector space, the sequence space \( I \), into another non-scalar discrete (or continuous) space \( S \). We call it a combinatory map \[81\] since the sequence space \( I \) is derived by a combinatory building principle (see also Sect. 3).

(ii) \( f(\cdot) \) maps a discrete (or continuous) non-scalar space \( S \) into the non-negative real numbers \( \mathbb{R}_+ \). It represents an example of a landscape or cost function. In particular, \( f(\cdot) \) is the fitness landscape assigning a fitness value \( f_k \) to a phenotype \( S_k \).

Finishing this section we consider environmental influences and indicate how one may generalize our approach to variable environments, \( E(t) \). Both, the unfolding of the genotype as well as the evaluation of the phenotype depend on the environment \( E \). In other words, the same genotype, \( I_k \) develops different phenotypes, say \( S_k, S'_k \) or \( S''_k \), in different environments, \( E, E' \) or \( E'' \). The same phenotype may have different fitness values under different environmental conditions. In principle, the ansatz for evolutionary dynamics presented here can be readily extended to handle situations in variable environments by the introduction of time dependent fitness values. Then we end up with the following equation which relates Darwinian fitness to genotypes:

\[ f_k(t) = f(S_k, E(t)) = f(\psi(I_k, E(t)), E(t)). \]

(8)

Incorporating time dependent fitness values into equations (1) and (3) we obtain a differential equation which is the basis for the deterministic description of Darwinian evolution in asexually replicating populations:

\[
\frac{dx_k}{dt} = x_k \left( Q_{kk} f_k(t) - \Phi(t) \right) + \sum_{j=1, j\neq k}^{n} Q_{kj} f_j(t) x_j , \quad k = 1, \ldots, n .
\]

(9)

Stochastic effects may be introduced into equation (9), for example, by means of a multidimensional master equation corresponding to a multivariate birth-and-death process with time dependent birth and death rates. Alternatively one may use Van Kampen’s size expansion of the master equation \[101\] and finally end up with a stochastic differential equation. Stochastic effects are then incorporated into the deterministic differential equation through terms

---

8The expression “landscape” is a generalization of the notion used in common-sense or geography for the representation of a three-dimensional relief on Earth as a mapping from two dimensions (longitude, latitude) into the real numbers (altitude).
like $\eta_k(x,t)\xi_k(t)$ which model fluctuations by a Wiener process whose amplitude depends on the variables of the deterministic solution. Separability of time scales is a prerequisite for the success of this approach: The environment driven changes in the functions $f_k(t)$ must be slow compared to the progress of the evolutionary process in order to allow for decoupling of external and intrinsic dynamics.

Needless to say, the mappings (8) encapsulate a great deal of complexity and there is no chance to find simple solutions. They are, nevertheless, suitable to discuss special simplified cases and they provide a proper reference for computer simulations. As said before, three experimentally accessible realizations of equation (9) are currently conceivable: (i) evolution of RNA molecules in vitro, (ii) life cycles and evolution of viral RNA (or DNA) in host cells, and (iii) metabolism and evolution of bacteria (under constant environmental conditions). In the following two sections we shall present a simplified model of (i) that allows to simulate evolution according to equation (9) using a realistic algorithm to compute RNA structures from sequences. Virus evolution (ii) can be modeled in principle provided enough data are available on the influence of mutations on viral life cycles. Quantity and quality of these data are rapidly improving now and we can expect full understanding of virus evolution at the molecular level within the forthcoming years. Although (iii) seems to be too complex by far for computer implementations we may expect fast progress in the near future: Information on complete DNA sequences is already available in a few cases and many more bacterial genomes will be sequenced soon. The current data are already used in the development of models for the genetic network controlling metabolism of prokaryotic cells [62]. Still, simulation of bacterial evolution based on such models will remain a great challenge for research in the future.

3 THE RNA MODEL

The phenotype in serial transfer or flow reactor experiments with RNA molecules is straightforwardly identified with the molecular structure of RNA [92]. Accordingly, the genotype-phenotype map relates RNA sequences with RNA structures. At the current state of the art our knowledge on RNA structures is far from being complete and hence prediction of RNA structures from known sequences is still a great challenge in bioinformatics and structural biology. The RNA case, however, is at least accessible by means of simplified but realistic models of sequence-structure mappings and thus contrasts the other, more complex phenotypes for which we have at best only pointwise genotype-phenotype information. The results of mathematical models and numerical computation on RNA optimization can be tested through comparison with the data from in vitro evolution experiments [5]. These data were complemented by results on RNA sequence-structure maps obtained from systematic studies based on site directed mutagenesis in RNA sequences (for example
the work on tRNAs [79]). Additional, highly valuable information comes from SELEX (systematic evolution of ligands by exponential amplification) experiments with RNA molecules aiming at the production of aptamers [27] as well as from design of ribozymes, which are specific catalysts on RNA basis (polyribonucleotide enzymes), through selection methods [3] (For an excellent and comprehensive review see [109]). Aptamers are optimized RNA molecules that bind specifically to predefined targets. Successful selection of optimal binders to almost all classes of biomolecules have been reported. Ribozymes were even obtained for reactions which have no counterpart in nature [78]. A recent and particularly interesting paper [84] reports the design of an RNA sequence, which adopts two conformations with different catalytic activities, specific RNA ligation and cleavage. Here we refrain from details, describe the current state of the art in the analysis of RNA sequence-structure maps, and mention only experimental results, which have direct implications for the theoretical concepts.

3.1 RNA PHENOTYPES

At present it is not yet possible to compute full three-dimensional RNA structures with sufficient reliability from sequences. A coarse-grained version of RNA structure, called secondary structure, however, is sufficiently simple in order to allow systematic investigations of genotype-phenotype maps (figure 2). Secondary structures, in addition, are not only convenient theoretical constructs but represent also relevant and experimentally verified intermediates in the folding of RNA sequences into three-dimensional objects [4]. Moreover, secondary structures are conserved in nature and they were used by biochemists for decades to interpret successfully the reactivities and other properties of RNA before three-dimensional structures became available. RNA secondary structures are understood best as a listings of Watson-Crick (AU, GC) and GU wobble base pairs, which are compatible with unknotted and pseudoknot-free two-dimensional graphs.\(^9\)

The simplest notion of an RNA genotype-phenotype map, and the one we shall adopt here, assigns the minimum free energy (mfe) structure which can be obtained through application of a suitable folding algorithm (see section 3.2) to the sequence under consideration. This assignment, apparently, makes use of the thermodynamic concept of structure in the limit of 0K. At nonzero temperatures we have to consider contributions from suboptimal foldings. The contributions of such configurations with energies higher than

\(^9\)RNA secondary structures can be represented by strings written in a short-hand notation using parentheses and dots. Parentheses correspond to bases combined in base pairs, dots represent single bases. The symbols for bases belonging together in pairs are interpreted unambiguously through reading them in the sense of mathematical notation, i.e. from outside to inside. For example, the string of a typical hairpin loop reads: \(\text{··}(((\text{··}))\text{··})\). A pseudoknot occurs when base pairs intercalate, for example in the secondary structure \(\text{··}(((\text{··}))\text{··})\), where we need two classes of symbols, parentheses and square brackets, for an unambiguous grouping of bases into pairs.
the mfe may be considered individually or handled collectively by choosing
the partition functions rather than the single mfe-structure as the phenotype.
This choice leads to a temperature dependent notion of phenotype. We are
thus dealing with a concept that allows for a straightforward response of the
phenotype to changes in an environmental parameter, the temperature. As
far as computational possibilities are concerned both, individual suboptimal
foldings [111, 114] and partition functions [45, 64], are accessible by efficient
algorithms based on dynamic programming.

The thermodynamic notion of structure, whether complemented through
suboptimal conformations or not, supposes the existence of an observation
window of infinite time for RNA folding. In reality, however, time is limited
and accessible structures are restricted by the necessity to fold sufficiently
fast. Then, the conformations formed are often different from the thermo-
dynamically most stable structures [71]. Such conformations are metastable
states and commonly addressed as kinetically controlled structures. Kinetic
folding has also been the subject of computations. Several computer programs
were designed to determine kinetic structures [41, 42, 61, 67–69, 94, 98]. These
algorithms are based on the concept of cooperative formation and melting of
double helices. Hence, they treat whole stacks as the units of structure which
form and open through all-or-none processes. A French group tried successfully
to integrate RNA folding into the general concept of stochastic chemical ki-
netics [7, 48, 49]. In a more recent paper an attempt was made to drag folding
down to the resolution of single base pair operations [28]. These operations
are closure and opening of base pairs as well as a shift move converting a base
pair into another allowed pairing of nucleotides. Kinetic folding, in particular
folding at single nucleotide resolution, introduces a new dimension into RNA
phenotypes: Not only thermodynamic stability but also the probability of for-
mation within a given time span determine the accessibility of a phenotype.
There is also a relevant third property, attainability through mutation [32,33],
which will be discussed in section 4.

How do the properties of RNA phenotypes change when the concept is ex-
tended from mfe-structures to suboptimal conformations, partition functions,
and kinetic folding? The answer in terms of biological concepts is straightfor-
ward: The mfe-structure regarded as a phenotype is relatively independent of
the environment and its response to changes is very limited. Then adaptation
is an almost exclusive property of populations which as a whole can cope with
variable environments through modifying and shifting genotype distributions
in sequence space. Suboptimal conformations or the partition function intro-
duce flexibility or “plasticity” in biological terms: The individual phenotype
can adjust to the environment by changing the distribution of conformations.
Considering RNA molecules in solution, variable environments may be visu-
ialized by changes in temperature, pH or ionic strength. Alternatively, binding
to other partners, for example small molecules, proteins or nucleic acids, may
also shift the conformational distribution and hence flexible phenotypes can
respond to the appearance of new molecules in their environments. Explicit
TABLE 1 Various strategies applied to study sequence-structure maps of RNA

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantage</th>
<th>Disadvantage</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematical model</td>
<td>Random graph theory</td>
<td>Analytical expressions</td>
<td>[81]</td>
</tr>
<tr>
<td>Exhaustive folding and enumeration</td>
<td>Folding algorithm</td>
<td>Exact results</td>
<td>[39, 40]</td>
</tr>
<tr>
<td></td>
<td>and handling of large</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>samples up to $10^{10}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical evaluation</td>
<td>Inverse folding or random</td>
<td>Applicability to longer sequences</td>
<td>[29, 88]</td>
</tr>
<tr>
<td></td>
<td>walks in sequence space</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulation of evolutionary dynamics</td>
<td>Chemical kinetics of</td>
<td>Evolutionary relevance</td>
<td>[32, 33]</td>
</tr>
<tr>
<td></td>
<td>replication and mutation</td>
<td></td>
<td>[47, 103]</td>
</tr>
</tbody>
</table>

Consideration of folding kinetics brings the time coordinate on the stage. It matters, whether a conformation can be adopted in sufficiently short time or not and whether a structure is formed with high or low probability. With respect to time we see even an (admittedly vague) analogy between RNA folding and development: Embryonic pattern formation or morphogenesis is also bound to occur within a sufficiently short time span. Otherwise the phenotype could not compete successfully in evolution.

In the following sections we shall adopt the simplest possible notion of phenotype, the mfe-structure. The more complex concepts discussed here can be incorporated straightforwardly into analysis of genotype-phenotype mappings and simulations of evolutionary optimization, although the higher computational efforts may be critical for the current possibilities.

### 3.2 SECONDARY STRUCTURES OF MINIMAL FREE ENERGIES

RNA secondary structures with minimum free energies are readily derived from sequences by means of fast algorithms based on dynamic programming [45, 74, 115, 116]. The mfe-structures, sometimes called (RNA) shapes for short, were studied in order to explore the regularities of sequence to structure mappings by means of four strategies (table 1): (i) mathematical modeling based on random graph theory [81], (ii) folding all sequences belonging to sequence space $I_\ell^X$ and computation through exhaustive enumeration [39, 40], (iii) computation through statistics of properly chosen samples
A recursion to calculate the numbers of acceptable RNA secondary structures, \( N_S(\ell) = S^{(\min[n_{lp}], \min[n_{st}])}_\ell \) [46]. A structure is acceptable if all its hairpin loops contain three or more nucleotides (loopsize: \( n_{lp} \geq 3 \)) and if it has no isolated base pairs (stacksize: \( n_{st} \geq 2 \)). The recursion \( m + 1 \Rightarrow m \) yields the desired results in the array \( \Psi_m \) and uses two auxiliary arrays with the elements \( \Phi_m \) and \( \Xi_m \) which represent the numbers of structures with or without a closing base pair \((1, m)\). One array, e.g. \( \Phi_m \), is dispensible but then the formula contains a double sum which is harder to interpret.

\[
\Xi_{m+1} = \Psi_m + \sum_{k=5}^{m-2} \Phi_k \cdot \Psi_{m-k-1}
\]
\[
\Phi_{m+1} = \left\lfloor \frac{(m-2)/2}{2} \right\rfloor \sum_{k=1}^{m-2k+1} \Xi_{m-2k+1}
\]
\[
\Psi_{m+1} = \Xi_{m+1} + \Phi_{m-1}
\]

Recursion: \( m + 1 \Rightarrow m \)

Initial conditions:
\[
\Psi_0 = \Psi_1 = \Psi_2 = \Psi_3 = \Psi_4 = \Psi_5 = \Psi_6 = 1
\]
\[
\Phi_0 = \Phi_1 = \Phi_2 = \Phi_3 = \Phi_4 = 0
\]
\[
\Xi_0 = \Xi_1 = \Xi_2 = \Xi_3 = \Xi_4 = \Xi_5 = \Xi_6 = \Xi_7 = 1
\]

Solution: \( S^{(3,2)}_\ell = \Psi_{m=\ell} \)

of sequences [29, 34, 88], and (iv) evaluation through evolutionary optimization [32, 33]. The following generic results were obtained:

(i) **More sequences than structures.** The numbers of acceptable secondary structures can be counted through combinatorial analysis of the assembly of conformations from elements by means of predefined rules [46, 105].

The calculations are done by means of the recursion shown in table 2.

---

A structure is considered acceptable if all hairpin loops contain three or more nucleotides and all stacks consist of at least two base pairs. Smaller loops are unstable because of high steric strain energies. Single base pairs are often unstable since the dominating stabilizing contribution comes from base pair stacking. Indeed, hairpin loops with one or two nucleotides are unknown in real structures and single base pairs occur only rarely. Despite
TABLE 3  Comparison of exhaustively folded sequence spaces [37,39,40,85,90]. The values are derived through exhaustive folding of all sequences of chain length \( \ell \) from a given alphabet. The parameters used were those which are currently implemented in the Vienna RNA Package [45,104]. The numbers refer to actually occurring minimum free energy structures without isolated base pairs which are directly comparable to the total numbers of acceptable structures \( N_S(\ell) = S^{(3,2)}_\ell \) (See table 2).

<table>
<thead>
<tr>
<th>( \ell )</th>
<th>Number of Sequences</th>
<th>Number of Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( 2^\ell )</td>
<td>( 4^\ell )</td>
</tr>
<tr>
<td>7</td>
<td>128</td>
<td>( 1.64 \times 10^4 )</td>
</tr>
<tr>
<td>8</td>
<td>256</td>
<td>( 6.55 \times 10^4 )</td>
</tr>
<tr>
<td>10</td>
<td>1,024</td>
<td>( 1.05 \times 10^6 )</td>
</tr>
<tr>
<td>12</td>
<td>4,096</td>
<td>( 1.68 \times 10^7 )</td>
</tr>
<tr>
<td>15</td>
<td>( 3.28 \times 10^4 )</td>
<td>( 1.07 \times 10^9 )</td>
</tr>
<tr>
<td>16</td>
<td>( 6.55 \times 10^4 )</td>
<td>( 4.29 \times 10^9 )</td>
</tr>
<tr>
<td>20</td>
<td>( 1.05 \times 10^6 )</td>
<td>( 1.10 \times 10^{12} )</td>
</tr>
<tr>
<td>25</td>
<td>( 3.36 \times 10^7 )</td>
<td>( 1.13 \times 10^{15} )</td>
</tr>
<tr>
<td>30</td>
<td>( 1.07 \times 10^9 )</td>
<td>( 1.15 \times 10^{18} )</td>
</tr>
</tbody>
</table>

For large chain lengths \( \ell \) the numbers \( N_S(\ell) \) are well approximated by the expression:

\[
N_S(\ell) \approx s(\ell) = 1.4848 \times \ell^{-3/2}(1.84892)^\ell.
\]

\( s(\ell) \) is an asymptotic upper limit for \( N_S(\ell) \) (For \( \ell = 30 \), for example, the deviation is 20.8\%, for \( \ell = 100 \) it is 6.0\%, for \( \ell = 300 \) it is smaller than 2.0\%, and for \( \ell = 1000 \) smaller than 0.65\%). Numbers computed from this expression (or the exact values) are many orders of magnitude smaller than \( 4^\ell \), and even orders of magnitude smaller than \( 2^\ell \), the cardinalities of sequence spaces built over four-letter and two-letter alphabets, respectively. Still we have only an upper limit for the number of shapes actually formed by folding all sequences of a given sequence space \( \mathcal{T}_\kappa^{\ell} \), which evidently obeys their high combinatorial probability minimum free energy conformations with single base pairs do not occur with GC-sequences at chain lengths \( \ell \leq 13 \). For longer sequences these conformations occur with increasing percentage: 5.7\% for \( \ell = 14 \), 12.8\% for \( \ell = 20 \), and 21.1\% for \( \ell = 30 \). The loopsize-restriction is less powerful in reducing the number of shapes than the neglect of structures with isolated base pairs: For chain length \( \ell = 30 \) we find \( 2.15 \times 10^{10} \) possible secondary structures, \( 2.41 \times 10^8 \) structures with loopsize \( n_{lp} \geq 3 \), and only 760,983 structures with loopsize \( n_{lp} \geq 3 \) and stacksize \( n_{st} \geq 2 \).

\[ |S_\kappa(\ell)| \leq N_S(\ell) \]. The cardinality of shape space, \( |S_\kappa(\ell)| \), can be obtained only by exhaustive folding and enumeration of mfe-structures.

As an example we consider binary GC-sequences of chain length \( \ell = 30 \). The number obtained from the recursion formula is \( N_S(30) = 760,983 \), whereas exhaustive enumeration of the shapes formed by binary GC-sequences yields \( |S_{GC}(30)| = 276,570 \) (For the parameters used in these computations see [104]). As many as 58,252 structures contain single base pairs resulting in 218,318 conformations representing acceptable mfe structures and being comparable to the computed number \( N_S(30) = 760,983 \) (table 3).\(^{11}\) This is only a fraction of 28.7\% of all acceptable structures, \( N_S(30) \). A comparison of \( |S_{GC}(30)| \) to the cardinality of sequence space, \( |I_{GC}(30)| = 1.07 \times 10^9 \), shows that the ratio of these numbers is indeed very small, \( |S_{GC}(30)|/|I_{GC}(30)| = 2.04 \times 10^{-4} \). In other words, the mean number of sequences forming the same structure is 4,901. In case of four-letter sequences the sequence to structure ratio would be much larger since we have \( |I_{AUGC}(30)| = 1.15 \times 10^{18} \) (see also table 4). Thus we are dealing with many more sequences than shapes. Clearly, the mapping from sequence space onto shape space is many to one and hence noninvertible.

(ii) **Few common and many rare shapes.** The distribution of the numbers of sequences forming the same shape, \( |S_k| \), is rather broad and strongly biased towards the rare-shape end. Analysis through exhaustive folding [39, 40] yielded a clear result independently of chain lengths \( \ell \) and size of alphabet (AUGC: \( \kappa = 4 \); GC: \( \kappa = 2 \)): There are relatively few common shapes and many rare ones. In the above mentioned example, GC-sequences of chain length \( \ell = 30 \) (GC\(_{30}\)), more than 93\% of all sequences fold into common shapes which are made up of only 10.4\% of all shapes. An increase in chain length causes these percentages to go up and down, respectively, and in the limit of long chains almost all sequences fold into a vanishingly small fraction of all shapes. It is worth to look at the GC\(_{30}\) shape space more closely [86]:\(^{12}\) The most frequent structures are formed by more than 1.5 million sequences which is about 0.15\% of sequence space. The shape of rank 10 (the tenth common structure) has still a pre-image of more than 1.2 million sequences. A glance at the rare frequency end is also illuminating: 12,362 shapes are formed by a single sequence only; 41,487 shapes by five or less sequences; the average number of sequences forming the same shape is 4,906, but 124,187 shapes, which more than 57\%, are formed by \( \leq 100 \) sequences.

---

\(^{11}\)The fraction of structures with isolated base pairs increases with the chain length \( \ell \). It starts in the GC-case with 5.7\% at chain length \( \ell = 14 \), reaches 12.8\% at \( \ell = 20 \) and 21.1\% at \( \ell = 30 \).

\(^{12}\)The numbers reported here are taken from the data of [39, 40] which were computed on the basis of a previous parameter set [50]. For comparison, the currently used parameters yield 161,756 shapes formed by \( \leq 100 \) sequences and 16,211 shapes formed by a single sequence only.
(iii) **Shape space covering.** Sequences forming common shapes are distributed (almost) randomly in sequence space. Accordingly, one need not search entire sequence space in order to find a sequence that folds into a given common shape. One can indeed show that is sufficient to screen a (high-dimensional) sphere around an arbitrarily chosen reference sequence in order to find (with probability one) at least one sequence for every common shape [88]. The radius of this shape space covering sphere, \( r_{\text{cov}}(\ell) \), can be estimated straightforwardly [85, 86]:

\[
 r_{\text{cov}}(\ell) = \min \left\{ h = 1, 2, \ldots, \ell \mid B_h(\ell, \kappa) \geq \frac{\kappa^\ell}{N_S(\ell)} \right\},
\]

where \( B_h \) is the number of sequences contained in a ball of radius \( h \) and can be easily obtained from the recursion

\[
 B_h(\ell, \kappa) = \sum_{i=1}^{h} b_i(\ell, \kappa); \quad b_i = b_{i-1} \cdot \frac{(\kappa - 1)(\ell + 1 - i)}{i}; \quad b_0 = 1.
\]

The covering radius for common shapes is much smaller than the radius of sequence space \((\ell/2)\). For example, it amounts to \( r_{\text{cov}} = 15 \) for AUGC-sequences of chain length \( \ell = 100 \) and thus one has to search only a fraction of sequence space \( I^{(100)}_4 \) that contains a \( 4.52 \times 10^{37} \)-th of all sequences in order to find at least one sequence for each of the common shapes.

(iv) **Common structures form extended neutral networks.** The pre-image in sequence space of a given shape \( S_j \) is the set of sequences \( M_j = \psi^{-1}(S_j) = \{ I_k | \psi(I_k) = S_j \} \). A set of sequences can be converted into a graph \( \mathcal{G} = ([v], [e]) \) in sequence space with \([v], \) and \([e], \) denoting the vertices and edges, respectively. The neutral network \( \mathcal{M}_j \) of \( S_j \) is constructed by identifying the sequences in \( M_j \) with the nodes and drawing edges between all nearest neighbors in the sequence space \( I^{(\ell)}_\kappa \) (these are the pairs of sequences with Hamming distance \( d_{ij}^h = 1 \)):

\[
 \mathcal{M}_j = \left( v[\mathcal{M}_j] = \{ I_k \mid I_k \in M_j \}, \quad e[\mathcal{M}_j] = \{(I_k, I_{k'}) | I_k, I_{k'} \in M_j \text{ and } d_{k,k'}^h = 1 \} \right).
\]

The question, how sequences belonging to a neutral network \( \mathcal{M}_j \) are distributed in sequence space was answered by means of random graph theory [81, 82]. The central quantity of this approach is the average degree of neutrality of a given network, \( \bar{\lambda}(\mathcal{M}_j) = \bar{\lambda}_j \). It is, in other words, the mean fraction of neutral neighbors of sequences belonging to the network:

\[
 \bar{\lambda}_j = \frac{\sum_{I_k \in \mathcal{M}_j} \lambda_k / | \mathcal{M}_j |}{| \mathcal{M}_j |}, \quad \text{where } \lambda_k \text{ is the number of nearest neighbor sequences of } I_k \text{ which form shape } S_j \text{ divided by the total number of nearest neighbors, } \ell \cdot (\kappa - 1).
\]

Neutral networks show a kind of percolation phenomenon. They are connected and span entire sequence space if \( \bar{\lambda}_j \) exceeds...
FIGURE 5  Classes of secondary structures with different distributions in sequence space. The three classes of structures sketched in the figure differ with respect to the ends of the stacking region. Class 0 structures contain a stack which cannot be extended by closing an additional base pair, class 1 structures can extend the stack on one end, class 2 structures on both ends.

A critical threshold value, whereas they are partitioned into components with one dominating giant part and many small “islands” when \( \bar{\lambda}_j \) is below threshold:

\[
\mathcal{M}_j = \begin{cases} 
\text{connected :} & \bar{\lambda}_j > (\bar{\lambda})_{cr} = 1 - \kappa^{-\frac{1}{\kappa-1}}, \\
\text{partitioned :} & \bar{\lambda}_j < (\bar{\lambda})_{cr} = 1 - \kappa^{-\frac{1}{\kappa-1}}, 
\end{cases}
\]

where \( \kappa \) is the number of digits in the alphabet of nucleotide bases (AUGC; \( \kappa = 4 \)). Connected areas on neutral networks are important in evolution since they define regions in sequence space which are accessible to populations through random drift [47].

The predictions on sequence-structure mappings of RNA shapes made by random graph theory were tested through exhaustive folding of entire sequence spaces [37, 39, 40]. In some cases we found deviations from generic behavior and these deviations could be explained by or derived from specific molecular structures. One particularly relevant and illustrative example was observed in the partitioning of neutral networks in the GC30 case. Random graph theory predicts that networks are either connected or their partition contains one largest “giant component”. Analysis of the sequence of components ordered with respect to sizes revealed, however, that there also networks with two or four dominant components of equal size, or with three components of size distributions 1:2:1. Most sequences of chain length \( \ell = 30 \) form shapes with
Evolution of Phenotypes

TABLE 4  Frequency of common shapes formed by AUGC-, GC-, and AU-sequences of chain length $\ell = 15$ as minimal free energy structures. Shapes are ranked according to their frequencies: The most frequent structure has rank no. 1, the next frequent one rank no. 2, etc. The open chain is not shown. It is the most frequent conformation in the AU- and AUGC-alphabet, and has rank 44 in the GC-shape space.

<table>
<thead>
<tr>
<th>Structure</th>
<th>AU-Alphabet</th>
<th>GC-Alphabet</th>
<th>AUGC-Alphabet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rank</td>
<td>Number of Sequences</td>
<td>Rank</td>
</tr>
<tr>
<td>(((((•••))))••)</td>
<td>2</td>
<td>748</td>
<td>5</td>
</tr>
<tr>
<td>((((••••))))••</td>
<td>3</td>
<td>640</td>
<td>21</td>
</tr>
<tr>
<td>((((•••••))))</td>
<td>4</td>
<td>618</td>
<td>4</td>
</tr>
<tr>
<td>•((((••••))))</td>
<td>5</td>
<td>472</td>
<td>22</td>
</tr>
<tr>
<td>••((((•••••))))</td>
<td>6</td>
<td>352</td>
<td>7</td>
</tr>
<tr>
<td>(((((•••••))))••)</td>
<td>7</td>
<td>264</td>
<td>10</td>
</tr>
<tr>
<td>((((••••))))</td>
<td>8</td>
<td>256</td>
<td>34</td>
</tr>
<tr>
<td>•(((((•••))))••)</td>
<td>9</td>
<td>176</td>
<td>39</td>
</tr>
<tr>
<td>•••((((•••))))</td>
<td>10</td>
<td>156</td>
<td>8</td>
</tr>
<tr>
<td>•••(((••••))))</td>
<td>11</td>
<td>80</td>
<td>11</td>
</tr>
<tr>
<td>•••(((((•••))))••)</td>
<td>12</td>
<td>68</td>
<td>13</td>
</tr>
<tr>
<td>•••(((••••))))</td>
<td>13</td>
<td>32</td>
<td>17</td>
</tr>
<tr>
<td>••••((((•••))))</td>
<td>14</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>••••(((((••••))))••)</td>
<td>15</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>(((((•••••))))••)</td>
<td>1</td>
<td>1 446</td>
<td>13</td>
</tr>
<tr>
<td>(((((•••))))••)</td>
<td>2</td>
<td>1 365</td>
<td>22</td>
</tr>
<tr>
<td>•••••(((•••))))</td>
<td>3</td>
<td>1 178</td>
<td>34</td>
</tr>
<tr>
<td>(((((•••••))))••)</td>
<td>4</td>
<td>618</td>
<td>4</td>
</tr>
<tr>
<td>••••••(((••••))))</td>
<td>2</td>
<td>748</td>
<td>5</td>
</tr>
<tr>
<td>(((••••••))))••</td>
<td>9</td>
<td>908</td>
<td>2</td>
</tr>
<tr>
<td>(((••••••))••)</td>
<td>18</td>
<td>668</td>
<td>3</td>
</tr>
<tr>
<td>••••••••((••••))))</td>
<td>26</td>
<td>501</td>
<td>4</td>
</tr>
<tr>
<td>•••••••••((•••••))))</td>
<td>35</td>
<td>380</td>
<td>5</td>
</tr>
<tr>
<td>••••••••••••((••••••••••••))))</td>
<td>25</td>
<td>505</td>
<td>6</td>
</tr>
</tbody>
</table>
one double-helical region. The four single stranded chains coming out from
the stack form a hairpin loop on one side and zero, one or two free ends on
the other side. Hairpin loops fall into two groups: (i) loops with \( n_{lp} = 3, 4 \)
and (ii) loops with five or more single bases. Loops of the former group cannot be
shortened by forming an additional base pair at the end of the stack since one
and two membered hairpin loops do not occur in real structures. In contrast,
five membered loops can be converted into a base pair and a triloop, six mem-
bered loops into a base pair and a tetraloop, etc. Similarly we find at the other
side of the stack: (i) no additional base pair can be formed if the number of
free ends is zero or one, but (ii) shapes with two free ends allow for elongation
of the stack provided the corresponding sequence requirements are fulfilled.
Combination of two elements at the two ends of the stack leads to three dif-
different classes of shapes (figure 5), which form neutral networks with different
sequence distributions in sequence space. Shapes with stacks containing two
category (i) ends of stacks (class \( 0 \)), these are tri- and tetraloops as well as
zero or one free ends, form generic neutral networks (connected or with one
largest component), shapes with one category (i) and one category (ii) end
(class \( 1 \)) form networks with two largest components, and shapes with two
category (ii) ends (class \( 2 \)), eventually, form those with three or four largest
components. Interpretation of this finding is straightforward: Generic neutral
networks (class \( 0 \)) show a distribution of sequences in sequence space which
is close to the binomial distribution (being fulfilled by the distribution of all
sequences of length \( \ell \) over an alphabet of size \( \kappa \) in sequence space):

\[
B(\ell, h, \kappa) = \binom{h}{\ell} (\kappa - 1)^h / \kappa^\ell ,
\]

where \( h \) is the Hamming distance between the reference sequence and the error
class under consideration (\( h \)). For binary sequences (\( \kappa = 2 \)) the distribution
\( B(\ell, h, 2) \) is symmetric in sequence space and the majority of all sequences is
found in the error class \( \ell/2 \) (or in the error classes (\( \ell \pm 1)/2 \), respectively). An
arbitrarily chosen most frequent binary sequence (50% G, 50% C) will form a
shape of class \( 1 \) with a reduced probability because 50% of these sequences
have complementary symbols (G, C) at the end of the stack and will sponta-
necessarily form the shape with the additional base pair. The largest probability
to form a class \( 1 \) shape is therefore displaced by some percentage \( \pm \delta \) from
the zone of highest frequency. As a matter of fact we find indeed two compo-
nents lying symmetrically with respect to the center of sequence space, one
with excess G (\( +\delta \)) and the other with excess C (\( -\delta \)). By the same token
we explain the occurrence of three or four largest components for shapes of
class \( 2 \): There are two labile category (ii) ends and the excess percentages of
G and C are superimposed independently yielding four components of equal
size (Displacements: \( +2\delta, +\delta - \delta = 0, -\delta + \delta = 0, -2\delta \)), two displaced ones
and two in the middle. Three components of sizes 1:2:1 originate from a four
component system through merging the two central components.
In the following, we compare a few of the most common shapes formed by AUGC-, GC-, and AU-sequences of chain length $\ell = 15$ (table 4) [89]. The limitation to such short chains is dictated by the cardinality of the four-letter sequence space (table 3) since handling of more sequences than a few billons is difficult and extremely time consuming. Nevertheless, the chains are already sufficiently long to allow for a certain variety of shapes: we find simple hairpins, e.g., $••(((••••)))$, and hairpins with internal loops or bulges, e.g., $••(((••••)))•$ or $(((••••••)))•$. No shapes with two hairpins were found at $\ell = 15$. They occur only with longer chains, $\ell \geq 16$, e.g., (((••))•(((••))))•.

In table 4 we compare the most common shapes in the three alphabets by their ranks. All structures listed in the table are simple hairpin loops. More complex structures are less frequent. The frequency of the open chain conformation (not shown in the table) is a measure of the importance of end effects. In the sequence spaces $I_{AUGC}^{(15)}$ and $I_{AU}^{(15)}$ the open chain is the most frequent conformation, formed by 42.43% or 88.12% of all sequences, respectively, and hence, end effects dominate. In $I_{GC}^{(15)}$ the open chain has rank no. 44 and is formed by only 0.98%. Because of the weakness of the AU base pair relative to GC, mfe-structures of AU-only sequences require long stacks in order to be sufficiently stable. Indeed, the fifteen structures of $S_{AU}^{(15)}$ are tri-, tetra-, and pentaloops attached to stacks of four, five, and six base pairs. Almost always these conformations are less frequent in the other two sequence spaces where smaller numbers of base pairs are already sufficient for stable conformations. Comparing $I_{AUGC}^{(15)}$ and $I_{GC}^{(15)}$ we see that the AUGC-alphabet sustains about 15% more shapes than the GC-alphabet, 152 versus 130 (table 3). As expected, the UGC-case takes a position between GC and AUGC. Although the ranking of shapes shows substantial differences in the different alphabets, the most frequent shapes in one alphabet correspond to common shapes in the other, an exception being the AU-case which does not sustain a sufficiently high number of stable conformations. The ranking of a shape according to its pre-image in sequence space is determined by two factors: In order to be frequent a shape has to have (i) sufficient thermodynamic stability (sequences that form shapes with positive energies are counted for the open chain), and (ii) high combinatorial probability on the sequence level. Clearly, both factors depend on the size and the nature of the alphabet.

In order to close this section we compare the results presented here with experimental data from a recent investigation. Schultes and Bartel [84] choose two RNA-conformations of chain length $\ell = 88$. The two folds have no base pair in common and the two parent sequences, LIG-P and HDV-P, represent two phylogenetically unrelated ribozymes with different catalytic activities, a synthetic specific RNA ligase [25] and a natural specific RNA cleaving ribozyme [77]. Then, the authors constructed one special RNA sequence, LIG-HDV, which is compatible with both secondary structures and thus can adopt both folds as stable or metastable conformations, respectively. The designed molecule did indeed show both catalytic activities, although the efficiencies
were smaller by four orders of magnitude than those of the parent molecules. The chimeric ribozyme sequence, LIG-HDV, was then optimized independently for both catalytic functions by mutation and selection. This means, two different optimization series were carried out which led into different directions of sequence space. Only four and two point mutations were required to yield sequences, LIG-4 and HDV-2, which recovered the full catalytic efficiency of the “parent ribozymes”, LIG-P and HDV-P, respectively. Still both optimized RNA molecules were more than Hamming distance 40 away from the parent sequences. For both RNA folds the authors were able to track neutral paths of constant structure and full ribozymic activity from the mutant to the parent, i.e. from LIG-4 to LIG-P and from HDV-2 to HDV-P. This elegant work provides direct proof for the existence of extended neutral networks of both ribozymes. In addition, it shows that the two networks approach each other very closely in the surrounding of LIG-HDV, and thus can be understood as an experimental example of the intersection theorem [81]. Moreover, the sequence LIG-HDV is also an illustrative example of an RNA molecule which switches between two conformations [28].

4 DARWINIAN EVOLUTION IN SILICO

In this section the RNA model will be used as an example of a genotype-phenotype map in computer simulations of evolutionary optimization of RNA shapes or structure related properties. At first the simulation has to be embedded in a physically relevant environment and we choose the flow reactor shown in figure 6 as an appropriate device. Such an experimental setup is known as “chemostat” in microbiology and used for continuous cultures of bacteria [56]. Apart from flow terms the chemical reaction mechanism contains replication and mutation steps and is described by equations (3) or (9) with $\Phi(t) = \sum_j a_j x_j(t)V/N$. It is implemented as a stochastic process based on the underlying master equation. Individual trajectories are computed by means of an algorithm conceived and analyzed by Daniel Gillespie [35, 36]. Under the constraints of the flow reactor the population size fluctuates, has an expectation value of $N$ and a standard deviation of $\sqrt{N}$. The replication rate constants are determined according to fitness criteria. In previous simulations [30,31] the kinetic constants were derived from molecular structures by some predefined and biophysically motivated rules. Error-free replication and mutation are parallel reaction channels whose relative frequencies are given by equation (4). The single digit accuracy of replication, $q$, corresponding to a mutation rate $p = 1 - q$ per site and generation, is an input parameter of the computations. Previous computer simulations confirmed three basic features of molecular evolution: (i) Population sizes of a few thousand molecules are sufficient for RNA optimization, (ii) stochastic effects dominate in the sense that the sequence of events recorded in one particular trajectory were
Evolution of Phenotypes

FIGURE 6  The flow reactor as a device for RNA structure optimization. RNA molecules with different shapes are produced through replication and mutation. New sequences obtained by mutation are folded into minimum free energy secondary structures. Replication rate constants are computed from structures by means of predefined rules (see text). For example, the replication rate is a function of the distance to a target structure which was chosen to be the clover-leaf shaped tRNA shown above (white shape) in the reactor. Input parameters of an evolution experiment in silico are: the population size $N$, the chain length $\ell$ of the RNA molecules as well as the mutation rate $p$.

never observed again in subsequent identical simulations,\textsuperscript{13} and (iii) sharp

\textsuperscript{13}By “two simulation experiments under identical conditions” we mean that everything was kept constant except the seeds for the random number generators.
error thresholds as predicted by the quasispecies concept were observed in computer runs with different mutation rates.

More recently, computer simulations of replication and mutation in the flow reactor were used to show that evolution on the neutral network of a tRNA-structure corresponds to a diffusion process in sequence space where the diffusion coefficient is proportional to the mutation rate \[47\]. In this simulation as well as in the computer experiments described below, replication rates were assumed to depend on the shape of the molecule independently of the sequence folding into it. Under this assumption the neutrality condition for sequences folding into the same structure, \(a_k = a(S_j) \forall I_k \in M_j\), is fulfilled. In particular, a function of the kind \(a(S_j; S_\tau) = (\alpha + d_{j\tau}/\ell)^{-1}\) was used, where \(\alpha\) is some constant, \(\ell\) the chain length of the RNA, and \(d_{j\tau}\) the distance between structure \(S_j\) and the target structure \(S_\tau\) \[33\]. Many measures of distance between structures are conceivable \[29\], a particularly simple one is the Hamming distance between the short-hand “parentheses notations” (Sect. 3) of the two shapes. For this goal shapes are understood as strings of the three symbols, ‘(’, ‘)’, and ‘•’. Specific features like the efficiency of optimization and the time required to reach a particular goal are, of course, influenced by the model assumptions and parameters. Generic results concerning the course of evolution, however, were found to be largely independent of the specific choices of kinetic parameters, fitness functions, and distance measures.

Optimization of RNA structures was studied through simulations of the evolution of a population in the flow reactor \[32, 33\]. Every optimization run had an initial period during which almost every mutation led to an increase in fitness. After this first phase the approach towards the target structure which happened to be a tRNA clover-leaf occurs in steps: Periods of fast decrease in the structure distance to target averaged over the whole population,

\[
\bar{d}_{\tau}(t) = \frac{1}{n_s} \sum_{j=1}^{n_s} p^s_j(t) d^{s}_{j\tau},
\]

are interrupted by long quasi-stationary phases or epochs of almost constant average fitness (figure 7). In equation (10), \(n_s\) denotes the number of shapes, \(p^s_j(t) = N^s_j(t)/N\) is the frequency of structure \(S_j\), and \(N^s_j(t)\) the number of individual molecules with structure \(S_j\). The course of the evolutionary optimization process was reconstructed through determination of a series of phenotypes leading from an initial shape to the target structure, called the relay series of the computer experiment. The relay series is a uniquely defined and uninterrupted sequence of shapes. It is retrieved through backtracking, that is in opposite direction from the final structure to the initial shape. The procedure starts by highlighting the final structure and traces it back during its uninterrupted presence in the flow reactor until the time of its first appearence. At this point we search for the parent shape from which it descended by mutation. Now we record time and structure, highlight the parent shape, and repeat the procedure. Recording further backwards yields
Evolution of Phenotypes

FIGURE 7  The recording of an RNA structure optimization experiment in the flow reactor. The computer experiment starts from a homogeneous initial population of 1000 RNA molecules with an arbitrarily chosen sequence folding into some initial shape. Fitness expressed as replication rate is computed as a function of the distance between current ($S_j$) and target structure ($S_\tau$), $d_{j\tau}$ (For details see text). The target structure was chosen to be the clover leaf of tRNA^Phe ($\ell = 76$). The mean distance to the target structure of the entire population, $d_\tau(t)$ in equation (10) and plotted against time (black curve). The time scale represents the “real time” of the simulation experiment in arbitrary units. The whole simulation comprises about $1.1 \times 10^7$ replications. A mutation rate of $p = 0.001$ per site and replication was applied. From this computer experiment a relay series of 42 shapes (or phenotypes) was reconstructed through backtracking the phenotypes which lead to the target structure (see text and figure 8). The six most important shapes are shown at the top of the figure. The relay series is indicated by the stepfunction (grey) which assigns equal height to every shape. Transitions between phenotypes fall into two classes: (i) continuous (examples marked by A) and (ii) discontinuous (B). A more or less well defined initial period of about one hundred time units is characterized by fast decrease in the distance to the target (a). The individual discontinuous transitions are classified as “shifts”, “flips”, and “double flips”, and marked by b to h. The “silent shift” at $t \approx 460$ is neutral with respect to distance to target. Discontinuous transitions lead to major changes in RNA shapes which are followed by cascades of minor fitness-improving steps.
FIGURE 8  The relay series of the *in silico* optimization experiment described in figure 7. For details see text. It is worth noticing that a given shape may appear twice or more often in a relay series. Examples are the shapes 29 ≡ 34 and 36 ≡ 38.
TABLE 5  Statistics of evolutionary trajectories. Different trajectories of in silico evolution towards a tRNA target were recorded for different values of the population size $N$ [107]. All other parameters and conditions were chosen as described in figure 7. The length of the relay series is shown as the number of relay steps ($n_{rl}$). In addition, we show also the number of discontinuous or major transitions ($n_{mt}$) and the mean structure distance between the population at the end of the fast adaptive initial phase and the target shape:

$$d_{in,\tau}^n = \sum_{j=1}^{N(t_{in})} d_{j,\tau}^n.$$  Herein $d_{j,\tau}^n$ is the structure distance between $S_j = \psi(I_j)$ and the target shape $S_\tau$, and $t_{in}$ the time at which the initial phase ends.

<table>
<thead>
<tr>
<th>Population Size</th>
<th>Number of Runs</th>
<th>Number of Relay Steps $n_{rl}$</th>
<th>Number of Transitions $n_{mt}$</th>
<th>Initial Phase $d_{in,\tau}^n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>10</td>
<td>120.1 ± 114.0</td>
<td>6.5 ± 1.7</td>
<td>17.6 ± 2.3</td>
</tr>
<tr>
<td>2000</td>
<td>13</td>
<td>66.3 ± 25.8</td>
<td>6.5 ± 1.7</td>
<td>18.5 ± 2.3</td>
</tr>
<tr>
<td>3000</td>
<td>12</td>
<td>41.9 ± 16.6</td>
<td>6.3 ± 2.2</td>
<td>17.6 ± 2.4</td>
</tr>
<tr>
<td>10000</td>
<td>17</td>
<td>37.8 ± 11.8</td>
<td>5.7 ± 1.3</td>
<td>16.5 ± 1.0</td>
</tr>
</tbody>
</table>

a series of shapes and times of first appearance which ultimately ends in the initial population.\(^{14}\) The full relay series of the computer experiment of figure 7 is shown in figure 8. It contains 42 shapes produced through $n_{rl} = 41$ consecutive transitions (Six characteristic structures along the series are shown on top of figure 7).

Transitions between two consecutive shapes in the relay series fall into two classes, A and B. Basis for this classification is the frequency of occurrence through mutations of the sequences from the reference neutral network (figure 9) which manifests itself also in the underlying structural change. Class A transitions occur frequently on mutation and involve mostly minor changes like closing and opening of a base pairs in the immediate neighborhood of a stack. Another example of a frequent transition is the opening of a stack of marginal stability, for example the terminal stack in the tRNA clover-leaf (the upper vertical stack in the secondary structure in figure 2): A mismatch in one base pair which is readily produced by a single point mutation is sufficient to open the stack. Class B transitions are rare events in the sense that they occur only with special sequences. They lead to major changes in structure. Such major rearrangements involve simultaneous displacement of

\(^{14}\)It is important to stress two facts about relay series: (i) The same shape may appear two or more times in a given relay series. Then, it was extinct between two consecutive appearances. (ii) A relay series is not a genealogy which is the full recording of parent-offspring relations a time-ordered series of genotypes.
FIGURE 9  **Statistics of shapes in the boundary of tRNA\textsubscript{phe}**. The basis of the statistics are 2199 sequences folding into the clover-leaf structure shown in figure 2. All their one-error mutants, 501,372 in number, were folded. A fraction of 28\% formed the same clover-leaf as the reference sequence and thus belonged to the neutral network. The remaining 358,525 sequences folded into 141,907 distinct shapes. Curve a is a log-log plot of the rank ordered frequency of occurrence (full line, right ordinate). The neighborhood frequency is plotted in curve b (dotted line, left ordinate). The dotted vertical line is meant to separate regions with different scaling: A region of frequent occurrence (left) is distinguished from the power-law distribution (right), which is typical for scaling according to Zipf’s law \[112\].

several base pairs (Different subtypes of class B transitions were characterized as “shifts”, “flips”, and “double flips” depending on the details of the structural change \[33\]). The majority of rearrangements recorded in RNA optimization experiments with population sizes of a few thousand molecules are class A transitions (Four of them are marked in figure 7). Class B transitions are less frequent. For example, seven major changes are identified among the 41 transitions of the relay series shown in figure 8.

Class A and class B transitions can be generalized in terms of neighborhood frequencies of neutral networks (figure 9):

(i) **Continuous transitions** (A). They represent minor structural changes and lead to structures which are globally frequent in the neighborhood of the neutral network of the initial shape.
(ii) **Discontinuous transitions (B).** They involve major structural changes
leading to globally rare and only locally frequent structures. Accordingly,
discontinuous transitions require special sequences that allow major struc-
tural changes to occur on single point mutations.
Typical simulations show an initial period \(0 \leq t \leq t_{in}\); marked a in figure 7) of
cascading discontinuous and continuous transitions followed by a stepwise opti-
mezation process with apparent regularities. Each epoch or quasi-stationary
phase of evolution ends with a discontinuous transition. Discontinuous tran-
sitions (b to h), however, occur only rarely within quiescent periods.\(^{15}\) Every
discontinuous transition is followed by a cascade of continuous transitions
which are accompanied by fitness increase. Then, the population approaches
the next plateau corresponding to an epoch of neutral evolution at approxi-
matelty constant fitness. Along the plateau the relay series shows neutral
mutations with respect to structure or fitness neutral class A transitions until
it reaches one of the special sequences from which a fitness-improving discon-
tinuous transition is locally frequent and hence attainable with sufficiently
high probability. Evolutionary optimization on landscapes with high degree
of neutrality proceeds on two time scales: Fast periods containing cascades
of adaptive changes are interrupted by long quasi-stationary epochs of neu-
tral evolution during which populations drift randomly on neutral networks
until they reach a neighborhood that is suitable for the next discontinuous
transition.

The analysis of the computer simulation experiments led to a novel notion
evolutionary nearness between phenotypes which is based on the concept of
neutral networks [33]. In order to explain nearness we consider a shape \(S_j\), its
pre-image \(M_j = \psi^{-1}(S_j)\), and the corresponding network \(\mathcal{M}_j\). The boundary
of the network, \(B_j = \text{bd}(M_j)\), is the set of sequences that can be reached from
\(M_j\) by a single mutation event but do not belong to \(M_j\). Folding the entire set
into shapes yields a distribution of phenotypes, \(\Sigma_j\), which is the image of \(B_j\)
in shape space. The error rates applied in the computer simulations reported
here (almost always) lead either to correct replication or single point mutations
and, hence, the boundary is the set of genotypes which are produced as one-
error mutants of the genotypes belonging to the network.\(^{16}\) A ranked frequency
distribution of phenotypes in the boundary of a neutral network shows, in
general, two clearly separable zones (figure 9): A relatively small number of
frequent phenotypes is contrasted by a large number of rare phenotypes. The
most common shapes are of comparable frequency and usually closely related
to the parent shape of the neutral network \((S_j)\). The distribution of the rare
phenotypes fulfils a power-law distribution, known as Zipf's law [112], which
implies that the log(frequency)/log(rank)-plot is a straight line. The results

\(^{15}\) There is one case (d) at time \(t \approx 460\) in the computer simulation of figure 7. A
discontinuous transition is observed inside an epoch. We called it “silent” since it does not
change the distance to target and is neutral with respect to fitness.

\(^{16}\) At higher error-rates it might useful to define two-error, three-error or, in general
n-error boundaries [33].
FIGURE 10  Variability in genotype space during punctuated evolution. Shown are the results of a simulation of RNA optimization towards a tRNA target (analogous to the run in figure 7) with population size $n = 3000$ and mutation rate $p = 0.001$ per site and replication. The figure contains two plots with different measures of genetic diversity, $d_P(t, \Delta t)$ and $d_C(t, \Delta t)$ with $\Delta t = 8000$ replications, against time, which is expressed as the total number of replications performed so far, and the trace (grey) of the underlying trajectory recording average distance from target. The upper plot contains the mean Hamming distance between the population ($d_P$; dotted line, right ordinate) at time $t$ and time $t + \Delta t$ and the lower one shows the Hamming distance between the mean sequences at the same moments ($d_C$; full line, left ordinate). The arrow indicates a remarkably sharp peak of $d_C(t, 8000)$ at the end of the second long plateau which reaches a Hamming distance of about 10. Every adaptive phase is accompanied by a drastic reduction in the genetic diversity while genetic variation increases during quasi-stationary epochs. The mutant cloud, whose average size is expressed by $d_P(t, \Delta t)$, expands fast during neutral evolution and reaches diameters up to Hamming distance 25 whereas the center of the cloud migrates only at a speed of Hamming distance 1 per 8000 replications.

are essentially the same for the two different distributions presented in figure 9: (i) the frequency of occurrence which counts the total number of sequences in the boundary that form the shape in question, and (ii) the neighborhood frequency which counts the number of neighborhoods where the shape occurs.
FIGURE 11  Spreading of a population in genotype space during a quasistationary epoch. The individual figures are snapshots of the genotype distribution at times corresponding to 1.2, 1.36, 1.6, 4.0, 6.4, 6.64, 6.696, and 6.84 × 10^6 replications. In order to visualize spreading, genotype distributions were transformed to principal axes and individual sequences were projected onto the plane spanned by the two largest eigenvectors. Along the series we observe an important and characteristic feature of population spreading in neutral evolution: The populations break up in smaller subclusters which diffuse radially away from the center of the distribution (See also the model on neutral evolution discussed in [13, 47]). Whenever an innovation with increase in fitness happens in one of the subclusters this subcluster takes over further development and all other subclusters die out.

The transitions to high frequency shapes in the boundary are the ones we called continuous (and, in other words, they occur readily on single point mutations). The threshold between frequent and rare shapes in the boundary can be defined intuitively: It occurs near the ranks where the linear range of the log/log-plot starts (see the straight line in figure 9). Transitions to shapes of low frequency in the boundary do not occur readily because they have sufficiently high probability only at certain special positions on the network $\mathcal{M}_j$ (which, for example, have to be found by the population through random drift). Accordingly, we called them discontinuous.

The data on the distribution of shapes in the boundary of neutral networks suggest to consider nearness in a statistical sense. A shape $S_k$ is (statistically) near $S_j$ when its frequency in the boundary $B_j$ is above threshold. Let $\rho(S_k; S_j) = \gamma(S_k, S_j)/|B_j|$ be the frequency of occurrence of $S_k$ in $B_j$ where $\gamma(S_k, S_j)$ is the number of Hamming distance one contacts between the two neutral networks and $|B_j|$ the cardinality of the boundary. Further, let $\varepsilon$
be a properly defined threshold value for the frequency in the boundary, then the set $\Psi_\varepsilon(S_j) = \{ S_k \in \Sigma_j | \rho(S_k; S_j) \geq \varepsilon \}$ defines the statistical neighbors of $S_j$ which are accessible through continuous transitions. It is important to note that $\rho(S_k; S_j)$ does not fulfill the conditions of a metric: In general, it is neither symmetric, $\rho(S_k; S_j) \neq \rho(S_j; S_k)$ nor does it necessarily fulfill the triangle inequality (although, of course, $\gamma(S_k, S_j) = \gamma(S_j, S_k)$ is always true).

In other words, the statement $S_a$ is statistically near $S_b$ does not imply that $S_b$ is near $S_a$. This paradox, however, is readily solved when one recalls that two networks may have (very) different sizes in sequence space. The larger network can occupy a fairly high percentage of the positions in the boundary whereas, at the same time, the smaller one is present only at low frequency in the boundary of the larger network. The proper mathematical context for accessibility of phenotypes as well as continuity and discontinuity in evolution is still being developed [10,11] but one can recognize already the usefulness of the concepts of statistical neighborhoods for simple as well as for general and highly complex genotype-phenotype mappings.

The time dependence of genetic diversity during evolution in silico is shown in figure 10. We apply two measures to visualize the diversity of genotypes in the population: (i) the mean Hamming distance within or between populations,

$$d_P(t, \Delta t) = \frac{\sum_{j=1}^{N(t)} \sum_{k=1}^{N(t+\Delta t)} d^h(I_j, I_k)}{N(t) \cdot N(t + \Delta t)}$$

and (ii) the Hamming distance between the mean nucleotide sequences at two different times $t$ and $t + \Delta t$,

$$d_C(t, \Delta t) = \sqrt{\sum_{k=1}^{4} \left( \sum_{j=A,U,G,C} \pi_{j,k}^a(t) \pi_{j,k}^b(t + \Delta t) \right)}.$$

The vector $\pi^k(t) = \{ \pi_{A,j}^k(t), \pi_{U,j}^k(t), \pi_{G,j}^k(t), \pi_{C,j}^k(t) \}$ is the square-normalized distribution of nucleotides at position $k$: $\pi_i^k(t) = \alpha_i^k(t) / \sqrt{\sum_{j=A,U,G,C} (\alpha_j^k(t))^2}$ with

\[17\] For examples of pairs neutral networks of RNA molecules lacking symmetry in the statistical neighborhood see [32]. We mention one representative case: a clover-leaf tRNA shape ($S_{\text{tRNA}}$) and a conformation with three hairpins which originates from the clover-leaf through opening of the terminal stack ($S_{\text{3hp}}$; the terminal stack is the upper vertical stack of the secondary structure in figure 2). The latter, $S_{\text{3hp}}$, is near $S_{\text{tRNA}}$ in a statistical sense, but the inverse is not true, $S_{\text{tRNA}}$ is not statistically near $S_{\text{3hp}}$. The interpretation of this asymmetry is straightforward. Let $\gamma(S_{\text{tRNA}}, S_{\text{3hp}})$ be the number of Hamming distance one contacts between the two networks. The sizes of the two networks are very different: $|M_{\text{tRNA}}| \ll |M_{\text{3hp}}|$, and hence, the boundaries too, $|B_{\text{tRNA}}| \ll |B_{\text{3hp}}|$, which leads to $\rho(S_{\text{tRNA}}; S_{\text{3hp}}) = \gamma(S_{\text{tRNA}}, S_{\text{3hp}})/|B_{\text{3hp}}| \ll \rho(S_{\text{3hp}}; S_{\text{tRNA}}) = \gamma(S_{\text{3hp}}, S_{\text{tRNA}})/|B_{\text{tRNA}}|$ and thus the statistical neighborhood relation does not commute, q.e.d.
\[ \sum_{j=A,U,G,C} \alpha_j^{(k)} = 1. \]

The former distance, \( d_P(t, \Delta t) \), describes, in essence, the spreading of the population in sequence space whereas the latter, \( d_C(t, \Delta t) \), is a measure for the migration of the center of the distribution. In figure 10 time is measured in terms of replications rather than in “real time” in order to come as close as possible to the number of generations: On generation corresponds to \( N \) replications on the average. We recognize an increase in genetic diversity during the quasi-stationary epochs of apparent constancy in shape space. The quantity \( d_P(t, \Delta t) \) is an appropriate measure of the average diameter of the mutant cloud. The size of the mutant cloud increases with time on the fitness plateaus and drops drastically when the population undergoes a discontinuous transition at the end of the epoch. The change in the mean nucleotide sequence of the population increases also during the quasi-stationary phase and saturates at values slightly above Hamming distance 1 per 8000 replications. At the end of every epoch we see a sharp or spike-like peak in \( d_C(t, \Delta t) \) which indicates a bottleneck in genotype space through which the population passes during discontinuous transitions. In order illustrate the spreading of populations we recorded the image of the population in genotype space on the fitness plateau between \( 1.2 \times 10^6 \) and \( 6.9 \times 10^6 \) replications (figure 11). After having passed the bottleneck of the previous discontinuous transition the population starts instantaneously to expand. At a time corresponding to \( 2.8 \times 10^6 \) replications the population breaks up into subpopulations which diverge further in sequence space. At \( 6.4 \times 10^6 \) replications an advantageous mutant appears in one of the subclusters, this subcluster takes over, and all other subclusters die out. The corresponding fast shift in the consensus sequence of the population gives rise to the spike recorded in \( d_C(t, \Delta t) \) (figure 10). The observation of random drift and partitioning into subclusters is in agreement with previous results [13, 47] and shows that the replication-mutation mechanism cannot sustain diffuse mutant clouds because the positions of descendants in sequence space are inevitably close to those of their parents. Although the picture of genotypic versus phenotypic evolution obtained from in silico simulation is much more detailed than the results recorded with bacterial populations [76], we see general agreement in the fact that, compared to adaptive periods, genomic evolution is at least equally fast or even faster during the phases of phenotypic stasis in evolution.

In order to study the population size dependence of evolutionary optimization we performed several computer runs and calculated the statistics of trajectories (table 5). The number of relay steps \( (n_{rl}) \) shows vast scatter but decreases considerably with increasing population size \( N \). This finding is easily interpreted: At larger population sizes the relay series contains fewer structures because individual shapes are less likely to die out and thus stay longer in the population. Interestingly, the number of major transitions \( (n_{mt}) \) shows smaller scatter and stays fairly constant within the investigated varia-

---

\(^{18}\)The delay time \( \Delta t \) was chosen to be identical in both quantities, \( d_P(t, \Delta t) \) and \( d_C(t, \Delta t) \). We remark that \( d_P(t, \Delta t) \) changes hardly within the time span considered: \( d_P(t, 8000) \) is almost identical with \( d_P(t, 0) \).
tion of population sizes. Similarly, the mean distance from the shape at the end of the initial adaptive period ($a$ in figure 7) to the target structure, $d_{\text{in}, \tau}$, does not change significantly with population size. It is obvious to suggest that the number of major transitions and the distance from target after the initial phase are more or less set by the target shape itself. The approach from a randomly chosen initial structure towards a given target is determined by the size and the distribution of its neutral network in sequence space.

5 DARWINIAN EVOLUTION AND INFORMATION

The mechanism of Darwinian evolution makes use of the powerful interplay of chance and necessity being identified with variation and selection, respectively [87]. Both processes can be traced down to the molecular level and studied in vitro by means of physical and chemical techniques as well as in silico by computer simulation. In order to analyze evolutionary optimization in full detail we introduce here a dynamical concept that starts out from chemical kinetics of replication and extends conventional population genetics (i) by visualizing mutation as a process in genotype or sequence space, and (ii) by introducing the phenotype as an integral part of the model. Central to this concept is a genotype-phenotype map which is used to derive phenotypic properties from genomic sequences and, in particular, to evaluate fitness parameters which are incorporated into the rate equations for the genotypes. Because of the enormous complexity of ordinary phenotypes such maps are not available yet except for the most simple evolutionary scenario consisting of RNA evolution in the test tube where the phenotype is tantamount to the molecular structure and its properties. At present, the only tractable case of such a mapping is the sequence-secondary structure map of RNA molecules. An RNA model based on this admittedly simple but, nevertheless, realistic mapping was used, for example, in computer simulations of evolutionary optimization and yielded, in essence, four results that are readily generalized to other, more complex biological systems.

(i) Evolution may show punctuation even under precisely constant environmental conditions like those encountered in a flow reactor. In other words, no external triggers are required for a stepwise course of optimization processes. Evolution occurs on two time scales: Adaptive processes dominated by selection are comparatively fast, and random drift on neutral networks is slow. Both, adaptive evolution and random drift contribute to the success of optimization processes. It was assumed for long time that random drift is a kind of unavoidable noise and has no positive effect on evolution. Diffusion on neutral networks, however, enables populations to escape from local fitness optima which otherwise would act as evolutionary traps. This result supports the view on the role of neutral evolution elucidated by Emile Zuckerkandl [113] in the context of the development of genetic regulatory networks: He addresses neutral and nonneutral mutations as a “creative mix” in evolution.
(ii) Accessibility of phenotypes determines the progress of evolution. A notion of nearness between phenotypes has been developed which accounts for the existence of neutral networks [33]. Nearness is defined in a statistical sense with respect to the frequency of occurrence of phenotypes in the one-error neighborhood of neutral networks. A phenotype which is frequently found in this neighborhood can be reached from almost every sequence of the network and the corresponding transitions were characterized as continuous because they occur almost instantaneously. Transitions to phenotypes which are infrequent in the neighborhood occur rarely and then only at special positions of the network. They were denoted as discontinuous. In other words, discontinuous transitions are locally frequent but globally improbable. Within the frame of the RNA model the frequencies of transitions are readily interpreted in terms of probabilities of structural changes caused by single point mutations. Discontinuous transitions and the major phenotypic changes they are associated with suggest to interpret them as the “real innovations” in Darwinian evolution.

(iii) Changes in genotypes may but need not be reflected by changes in the phenotypes and thus the relative pace of genomic and organismic evolution is a central issue in biology. Computer simulations provide direct insight into this problem. The dispersion of genotypes in the population varies strongly and systematically during evolution. Strong increase of genomic diversity is observed during the diffusion on a neutral network. Simultaneously with the spread, the population is split into smaller clusters of sequences which represent individual clones. At the same time the center of the population in sequence space shows only small drift. A discontinuous transition is commonly manifested by a dramatic drop in the diversity of genotypes and a “jump” in the center of the population. Discontinuous transitions may be interpreted as “bottlenecks” in genomic evolution. The population becomes almost uniform during the passage and then spreads again on the next neutral network. The large shifts in genotype space observed with the population centers mean that the discontinuous transition started out from one particular clone which becomes dominant and then represents the new center on the population. It is straightforward to compare differences between evolution of genotypes and phenotypes in terms of changes per generation: Computer simulations show that genomic evolution speeds up through spreading of populations on neutral networks whereas phenotypic evolution measured in terms of fitness or distance to target is slow or practically zero as seen from the almost constant fitness plateaus. Major changes in phenotypes initiated by discontinuous transitions are accompanied by a drop in genetic diversity. Similar inverse relations between genomic and phenotypic change were found in the analysis of long time evolution experiments with bacteria [76].

(iv) How does the population size influence the evolutionary scenario described here? First of all, the larger the population size, the smaller is the number of individual steps in the relay series. Interestingly, the occurrence of major transitions as well as their numbers were found to be quite insensitive to variations
of the population size. Punctuated evolution was found, after all, also with bacterial populations whose size was in the order of $5 \times 10^6 < N < 5 \times 10^8$ cells or in experimental setups with up to $10^{15}$ RNA molecules. Apparently, the transition from stochasticity to deterministic behavior occurs at much larger population sizes as far as these phenomena are concerned. Because of the enormously high numbers of genotypes evolution remains inherently stochastic even at the largest population sizes that can be realized. The same seems to be true for evolution on neutral networks. Clearly, larger populations would stay longer connected and break up in subclusters later (or not at all). The other qualitative features discussed here, in particular the time dependence of $d_P(t, \Delta t)$ and $d_C(t, \Delta t)$ shown in figure 10, would remain largely uneffected by a change in population size.

(v) The landscape concept, originally introduced by Sewall Wright [110] as a metaphor into evolutionary biology, was put on firm scientific grounds when applied to biopolymers, in particular proteins and RNA molecules. Molecular biophysics revealed the basis of neutrality and neutral evolution which population geneticists could deduce only indirectly from comparisons of sequence data in contemporary organisms. The RNA model is currently based on shapes defined as minimum free energy structures. The concept, however, is very flexible because additional structural features can be taken into account readily. Such features are, for example, the consideration of suboptimal conformations within reach from the ground state at room temperature or the kinetics of the folding process. In addition, consideration of tertiary interactions in RNA molecules will lead to truly three-dimensional structures. We can indeed expect a great variety of new phenomena which wait to be detected in such extended molecular models. These new regularities will help to illustrate and understand otherwise too complex to analyze peculiarities of macroscopic evolution.

Finally, we address the question of genetic information and how it is created through evolution in Darwinian systems. The principle of variation and selection is a special case of self-organization and thus requires self-enhancement and non-equilibrium conditions. In biology and in assays which mimic biological evolution in vitro, self-enhancement is tantamount to multiplication. Furthermore, it is necessary to produce variants, on which selection can act, and to have a storage device which keeps track of the past in the sense of inheritance. All these requirements are already fulfilled with RNA molecules replicating in the test tube and, indeed, the origin of genetic information can be visualized within the concept of the molecular quasispecies [6, 21, 22, 24]. A population gains information in the course of the selection process since the selected molecules carry a kind of indirect “image” of their environment. The higher the fitness, the better is this image, since it allows for better exploitation of the resources.

The ultimate basis of information in biology is interaction between molecules or molecular recognition. This recognition is improved during selection. Creation of information, however, requires more than just increase in
strength and specificity of interactions, it requires the capacity of storage of past experience and retrieval in the sense of information processing. Although recognition is not restricted to a certain class of molecules, information processing is dependent on more stringent requirements that are fulfilled only by molecules which are both, sufficiently stable and able to act as templates in a copying process. Only a copying mechanism guarantees that replication errors, once they have occurred, can be transmitted to future generations and thus subjected to selection. One and the only class of molecules which are presently known to be suitable for this purpose are the nucleic acids. No wonder that evolution of molecules and its applications to biotechnology [27, 38, 97, 106] were more or less restricted to the use of RNA and DNA. Design of proteins or organic molecules by means of selection techniques require either translation, which couples protein design to nucleic acid evolution, or direct intervention by the experimentalist through screening techniques.

How do neutral networks and explicit consideration of phenotypes modify or change the conventional view of the origin of genetic information? The answer is not straightforward. First, the often spread selectionist’s view saying that (almost) “every adjustment to the environment is possible and can be achieved through an eventually very large number of infinitesimally small steps” is not true. The set of attainable conformations is usually substantially smaller than the set of all (possible) phenotypes, and steps need not be small. At the current state of the art, estimates on the accessible fraction of the “universe of possible phenotypes” are highly uncertain and we have to wait for more information before we can give a decisive answer. However, it is certainly possible to find examples demonstrating lack of accessibility: So far all attempts failed to obtain a tRNA-like shape through evolutionary optimization of RNA molecules built over a two-letter GC-alphabet, although it is straightforward to show that such molecules do exist and are stable (See, for example [33], p.513). Second, the sequences forming a neutral network for the mfe-structure commonly differ with respect to suboptimal conformations or folding properties. Since these properties may contribute to fitness as well, the migration of the population on the neutral network corresponding to the ground state conformation may in fact follow a (small) fitness gradient. In the case of such flexible phenotypes “neutral evolution” is not strictly neutral any more and may lead to structures which receive more complex properties through the appropriate suboptimal conformations and/or suitable folding patterns. Then, “guided drift” builds upon the properties that are not encoded in the conformation of the ground state, and thus diffusion on networks (which are strictly neutral only with respect to the mfe-structure) may also generate genetic information. Further investigations on RNA evolution of and with adaptable phenotypes will shed more light on the underlying mechanisms.
Acknowledgements

The work on the molecular quasispecies is the results of a cooperation with Manfred Eigen and John S. McCaskill then at the Max-Planck-Institute of Biophysical Chemistry in Göttingen, Germany and Karl Sigmund from the Institute of Mathematics at the University of Vienna, Austria. The RNA model was developed during long time research of our group at Vienna University. Many discussions with Manfred Eigen, Christoph Flamm, Walter Fontana, Ivo Hofacker, John McCaskill, Karl Sigmund, and Peter Stadler are gratefully acknowledged. Andreas Wernitznig kindly provided additional computer plots on in silico evolution experiments in the flow reactor. Financial support of the work presented here was provided by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung (Projects P-13 093, and P-13 887), by the Jubiläumsfonds der Österreichischen Nationalbank (Project 7813), by the Commission of the European Union (Project PL-97 0189), and by the Santa Fe Institute.

REFERENCES


[77] Perrotta, A. T. and M. D. Been, “A Toggle Duplex in Hepatitis Delta Virus Self-Cleaving RNA that Stabilizes an Inactive and a


