Molecular Insights into Evolution of Phenotypes

Peter Schuster

SFI WORKING PAPER: 2000-02-013

SANTA FE INSTITUTE

Molecular Insights into Evolution of Phenotypes

Peter Schuster
Institut für Theoretische Chemie und Molekulare
Strukturbiologie der Universität Wien
Währingerstraße 17, A-1090 Wien, Austria and
Santa Fe Institute, Santa Fe, NM 87 501, USA
pks@tbi.univie.ac.at

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 only through empirical parameters. 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 *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 *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 stucture to target, called relay series. We give a novel definition of continuity in evolution which identifies discontinuities as major changes in molecular phenotypes.

Evolutionary Dynamics — Exploring the Interplay of Accident, Selection, Neutrality, and Function
Edited by J. P. Crutchfield and P. Schuster, Oxford Univ. Press

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.[1] Recombination occurring obligatorily in sexually reproducing populations is another process contributing to the maintainance 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 targets for 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 [60]. 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 [82] pioneered experiments on evolution of RNA molecules in the test-tube (figure 1). A sample of RNA of the bacteriophage $Q\beta$ was transferred into a solution containing an RNA replicase and the

[1] The genotype is understood as the polynucleotide sequence which carries the genetic information to build the organism. The polynucleotide is commonly DNA, or RNA in the case of several families of viruses and viroids. The phenotype is the entity that carries all properties which are required to enter the reproductive phase. For higher forms of life the phenotypes are the adult organisms, for prokaryotes is the bacterial cell or the virus particle. The phenotype thus determines fitness which is commonly understood in evolutionary biology as the number of fertile descendants transmitted into the next generation.

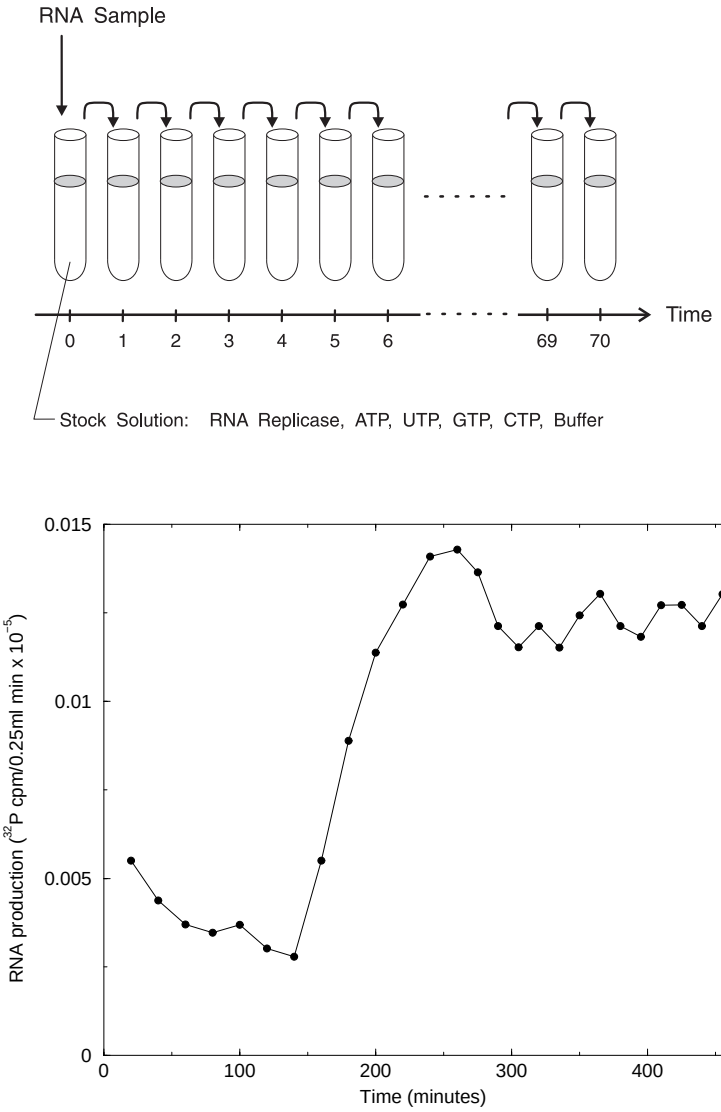

FIGURE 1 **Stepwise increase in the rate of RNA production.** The upper part shows the technique of serial transfer applied to evolution of RNA molecules in the test tube. The material consumed in the synthesis is replaced by transfer of a small sample into a new test tube with fresh stock solution. 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 [61].

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 hundered times. In suche 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 *a priori* that genotype and phenotype are sufficiently distinct in order to lead to *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 structure 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. 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 conformtaions may play a role also for rather small RNA molecules [26]. For longer RNA sequences the discrepancy between most stable and kinetically favored metastable structures is well established [66]. 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 conformations which resides in the valley under consideration. These conformations are accessible within the (temperature dependent) time window of observations. Modified Boltzmann ensembles corresponding to such subsets are good candidates for an elaborate notion of biopolymer structure.

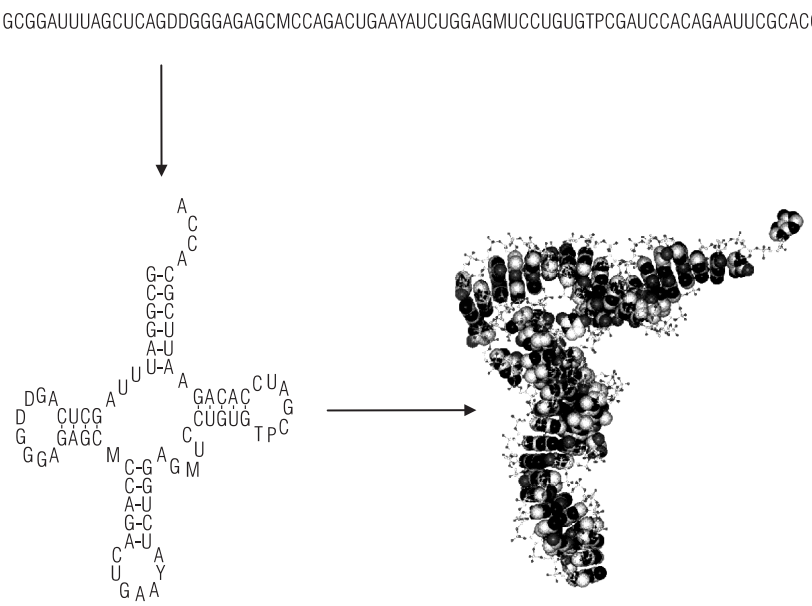

FIGURE 2 **Folding of RNA sequences into structures.** The folding is performed in two steps from the sequence to the secondary structure and from the secondary structure to the full spatial structure. The example shown is the transfer RNA molecule $tRNA^{phe}$. Both steps occur under the condition of 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.

An interesting feature of Spiegelman's RNA evolution and other *in vitro* evolution experiments is stepwise increase in fitness or other quantities used to monitor optimization. Punctuation is observed even under controlled constant conditions (figure 1). Epochal evolution [92] is not restricted to evolution of molecules in the test tube: It has been observed also with bacterial cultures under the constant conditions of precisely controlled serial transfer [24] as well as in evolution experiments *in silico* mimicking replication and mutation in a flow reactor [29–31, 45, 93]. A straightforward (but rather trivial) interpretation of the phenomenon says that the population waits for some rare event during such quasi-stationary periods. Basic questions concern the nature of the rare event and the strategy applied by the population in its search for an infrequent incident. We shall try to give answers which are compatible with the now well established neutral [52] or nearly neutral [69] 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 [16]. An elegant serial transfer experiment with *Escherichia coli* bacteria was carried out by Richard Lenski and coworkers [24, 53, 70]. Populations of 5×10^8 cells were diluted 1:100 every day and recorded for about three years leading to about 10 000 generations and an average generation time of 3.6 hours.[2] 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 an not continuously as one might have expected [24]. 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 [70]: Phenotypic evolution is fast in the initial phase and slows down during saturation. Evolution of the genome, on the other hand, behaves differently: it speeds 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, united the previously conflicting issues of Darwinian evolution and Mendelian genetics in an elegant and straightforward way. Evolution is considered as an optimization process on 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

[2] More precisely the bacteria in the solution multiplied substantially faster at the beginning of a transfer period and slowed down later when the nutirent fluid became exhausted.

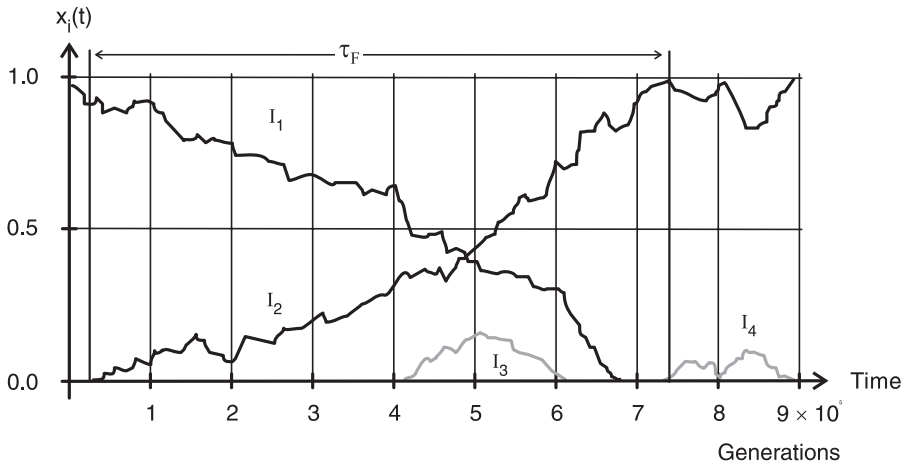

FIGURE 3 **Evolution of genotype frequencies in asexual populations.** The sketch shows typical solutions curves representing relative concentrations or frequencies $x_i(t)$ of a population modeled by equation (1). New variants are formed by rare mutation events. Depending on replication rates relative to the mean value the frequencies will increase ($a_i > \bar{a}$), decrease ($a_i < \bar{a}$) or, in the neutral case, drift randomly ($a_i \approx \bar{a}$). Stochastic theory shows that fixation of mutants occurs also in neutral evolution: According to Kimura's theory [52] the mean time from the appearence of a mutant, $x_m(0) = 1/N$, until its fixation in the population, $x_m(\tau_F) \approx 1$, is $< \tau_F > = 2N$.

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 populations. 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

[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.

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* [23] 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:

$$\frac{dx_k}{dt} \;=\; x_k \Big(e_k \;-\; \Phi(t) \Big) \,, \;\; k = 1, \ldots, n \;. \tag{1}$$

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 [19,49,54,55,85]. 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 [65], 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 [49].

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 [91] was applied to derive stationary solutions of the stochastic selection problem [55]. 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* [51,52]. 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, $< k >$, is measured as the number of mutant substitutions per generation time and can be expressed by

$$< k > = \ N \cdot v \cdot u(N,s,p) \ = \ N \cdot v \cdot \ \frac{1-\exp(-2Ns\,p)}{1-\exp(-2Ns)} \ , \qquad (2)$$

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,[4] 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

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

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 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

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

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 advatage: $\lim_{s\to\infty} < k >= N \cdot v$. 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. We should keep in mind, nevertheless, that the upper limit of $< k >$ is highly unrealistic because it requires to maintain large increases in fitness through mutation, which do not occur over a sequence of many consecutive mutants under normal conditions (see, however, the initial period of *in silico* evolution of RNA molecules in Sect. 4).

It is also worth considering the mean time for fixation of neutral and advantageous variants, τ_F. The solution of the deterministic selection equation for two genotypes,[5] $x_2 = x$, $x_1 = 1 - x$ and $f_0 = 1$, is readily obtained in analytical form:

$$x(t) \;=\; \frac{x_0}{x_0 \,+\, (1 - x_0) \cdot \exp(-s\,t)} \;.$$

From this equation we compute τ_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 [20]. 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 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 deterministic equation (1) is readily extended to handle the frequent mutation

[5] Thereby we mean equation (1) for $n = 2$.

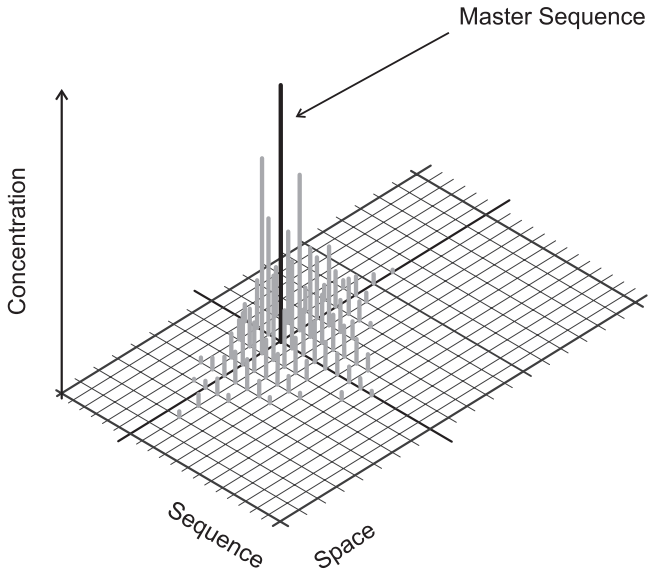

FIGURE 4 **A quasispecies-type mutant distribution around a master sequence.** The quasispecies is an ordered distribution of polynucleotide sequences (RNA or DNA) in sequence space $\mathcal{I}_\kappa^\ell$. A fittest genotype or master sequence I_m, which is present at highest concentration, is surrounded 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) [20, 23], the mutant distribution described by the largest eigenvector of the matrix $W = \{W_{ij} = Q_{ij} \cdot a_j; i, j = 1, \ldots, n\}$ [48, 67, 86, 90] (Its diagonal elements are approximations for fitness values, I_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^ℓ) exceeds the number of individuals by many orders of magnitude even in the largest populations. The actual behavior is determined by incorrect replication and random drift: populations migrate through sequence space.

scenario and we obtain the replication-mutation equation:

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

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.[6] 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* [22, 23], 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, $g : \{\mathcal{I}; d_{ij}^h\} \Rightarrow \mathbb{R}^1$ (For recent reviews see [77, 83]). Herein the sequence space space is denoted by $\mathcal{I}$ and the distance between two sequences by d_{ij}^h. Several classes of model landscapes were either invented, like for example the Nk-model [50], or adopted from physical models of spin-glasses [89] 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* [22, 23]. At the critical point an abrupt change in evolutionary dynamics is observed which reminds of a phase transition [56] (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).

[6] The symbol $I_\cdot$ is used here for genotypes or polynucleotide sequences (DNA, RNA) as well as $\mathcal{I}$ for the space of genotypes, called sequence space [42] in order to point out that the sequences are the carriers of biological information. In particular, $\mathcal{I}_\kappa^\ell$ is the sequence space of sequences of chain length ℓ over an alphabet of size κ (**GC**: $\kappa = 2$; **AUGC**: $\kappa = 4$). $\mathcal{I}_\kappa^\ell$ carries κ^ℓ different sequences. The Hamming distance [41] of two sequences I_i and I_j is denoted by d_{ij}^h 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 be or become unstable: (i) The steady state can never be reached because the largest population sizes that can be realized are too small to be an approximation to the infinite population limit, and (ii) random drift in the sense of neutral evolution (which is not addressed by deterministic equations) may lead to further onset of selection after epochs of stasis.

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 or shape space, are enormous, even for moderate chain length, and hence too large for any direct assignment of empirical parameter. 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 ($\mathcal{I}_\kappa^\ell$).[6] 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, ℓ, q and d_{kj}^h:

$$Q_{kj} \;=\; q^\ell \left(\frac{1-q}{q}\right)^{d_{kj}^h} \quad . \tag{4}$$

Herein ℓ 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}^h, 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 [20]. First, mutational backflow (the sum in the right part of equation 3) is neglected and second, the stationary frequency of the master genotype is computed to $\bar{x}_m = (\sigma_m Q_{mm} - 1) \;/\; (\sigma_m - 1)$, wherein $\sigma_m = a_m/\bar{a}_{k\neq m}$ 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, $p_{\max} = 1 - q_{\min}$) at constant chain length and one for the maximal chain length at constant mutation rate,

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

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 [22]. 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 [97] as well as to neutral evolution where stationarity refers to time independent distributions of phenotypes rather than genotypes [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 1.6×10^{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 $a_{k\neq m} = 1$ [86]. 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 [20, 22, 23] (For further work on replication and mutation on model landscapes see [1, 2, 7, 8, 58, 81, 89]).

An analysis of stochastic effects in replication-mutation systems based on multitype branching processes [11] 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 [67] 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 [19, 49, 54, 55, 85] and applied the same constraints as discussed for the selection equation, the Moran model [65] 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 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 [4, 75]. 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 a

population evolves at maximal mutation rate. This was indeed found to be the case with lytic RNA viruses [17] where the observed mutation rate p increases linearly with the reciprocal chain length ℓ^{-1}. A straightfoward 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 [16, 18] 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 [18]). In addition, the quasispecies concept turned out to be useful for the description of the evolution of RNA virua populations [13, 14] and provided hints for the developement of novel antiviral strategies [15].

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) [29]. 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 $\mathcal{S}$ with some (hypothetical) measure of distance between phenotypes, d_{ij}^s:

$$\psi : \{\mathcal{I}; d_{ij}^h\} \Rightarrow \{\mathcal{S}; d_{ij}^s\} . \tag{6}$$

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 acuuracy, $f_k \approx a_k \cdot Q_{kk}$, and can been seen as the result of a mapping f from the phenotypes into the nonnegative real numbers:

$$f : \{\mathcal{S}; d_{ij}^s\} \Rightarrow \mathbb{R}_+ . \tag{7}$$

The map (7) evaluates the phenotype and returns its fitness value. In summary, we obtain fitness values from the genotype through the function: $f(S_k) = f\big(\psi(I_k)\big) = f_k \approx a_k \cdot Q_{kk}$. The mapping $\psi(.)$, 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, $f(.)$ 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) $\psi(.)$ maps a discrete vector space, the sequence space $\mathcal{I}$, into another non-scalar discrete (or continuous) space $\mathcal{S}$. We call it a combinatory map [73] since the sequence space $\mathcal{I}$ is derived by a combinatory building principle (see also Sect. 3).
(ii) $f(.)$ maps a discrete (or continuous) non-scalar space $\mathcal{S}$ into the nonnegative real numbers $\mathbb{R}_+$. It represents an example of a landscape, in particular, it is the fitness landscape assigning a fitness value f_k to a phenotype S_k.[8]

Finishing this section we consider environmental influences and indicate how one may generalize our approach to variable environments, $\mathcal{E}(t)$. Both, the unfolding of the genotype as well as the evaluation of the phenotype depend on the environment $\mathcal{E}$. In other words, the same genotype, I_k develops different phenotypes, say S_k, S'_k or S''_k, in different environments, $\mathcal{E}$, $\mathcal{E}'$ or $\mathcal{E}''$. 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 with genotypes:

$$f_k(t) \;=\; f\Big(S_k, \mathcal{E}(t)\Big) \;=\; f\Big(\psi\big(I_k, \mathcal{E}(t)\big), \mathcal{E}(t)\Big) \;. \tag{8}$$

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:

$$\frac{dx_k}{dt} \;=\; x_k \Big(Q_{kk} f_k(t) \;-\; \Phi(t)\Big) \;+\; \sum_{j=1, j\neq k}^{n} Q_{kj} f_j(t)\, x_j \;, \quad k = 1, \ldots, N \;. \tag{9}$$

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 and finally end up with a stochastic differential equation. Stochastic effects are then incorporated into the deterministic differential equation through terms like $\eta_k(\mathbf{x}, t)\xi_k(t)$ 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

[8] The 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).

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. 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 metabolism of prokaryotic cells. Still, simulation of bacterial evolution based on such models will remain a great challenge for future research.

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 [82]. 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 [4]. 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 [71]). Additional information comes from SELEX experiments with RNA molecules aiming at the production of aptamers [25]. Aptamers are RNA molecules that bind optimally to predefined targets. Successful selection of optimal binders to alost all classes of biomolecules have been reported. Here

we refrain from details and describe only the current state of the art in the analysis of model for RNA sequence-structure maps.

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 a convenient theoretical constructs but respresent also a relevant and experimentally verified intermediates in the folding of RNA sequences into three-dimensional objects [3]. 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 stuctures 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 0 K. At nonzero temperatures we have to consider contributions from suboptimal foldings. The contributions of such configurations with energies higher than 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 environmetal parameter, the temperature. As far as computational possibilities are concerned both, individual suboptimal foldings [99,102] and partition functions [59], 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-

[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: $\cdot\cdot((((\cdot\cdot\cdot))))\,.$ A pseudoknot occurs when base pairs intercalate, for example in the secondary structure $\cdot\cdot(((\cdot\cdot[[\cdot\cdot\cdot)))\cdot]]\cdot$, where we need two classes of symbols, parentheses and square brackets, for an unambiguous grouping of bases into pairs.

dynamically most stable structures [66]. 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 [39, 40, 57, 62–64, 84, 88]. 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 sucessfully to intergrate RNA folding into the general concept of stochastic chemical kinetics [6, 46, 47]. In a more recent paper an attempt was made to drag folding down to the resolution of single base pair operations [26]. 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 formation within a given time span determine the accessibilty of a phenotype. There is also a relevant third property, attainability through mutation [30,31], which will be discussed in section 4.

How do the properties of RNA phenotypes change when the concept is extended from mfe-structures to suboptimal conformations, partition functions, and kinetic folding ? The answer in terms of biological concepts is straightforward: 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 introduce 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 visualized 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 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 probabilty. With respect to time we see even an (admittedly vage) 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.

TABLE 1 Various strategies applied to study sequence-structure maps of RNA

	Method	Advantage	Disadvantage	Ref.
Mathematical model	Random graph theory	Analytical expressions	Limited validity of model assumptions	[73]
Exhaustive folding and enumeration	Folding algorithm and handling of large samples ($> 10^9$ objects)	Exact results	Limited to short chains: **GC**, $\ell \leq 30$ **AUGC**, $\ell \leq 16$	[37, 38]
Statistical evaluation	Inverse folding or random walks in sequence space	Applicability to longer sequences	Limited accuracy due to statistics	[27, 79]
Simulation of evolutionary dynamics	Chemical kinetics of replication and mutation	Evolutionary relevance	Restriction to small parts of sequence space	[30, 31] [45, 93]

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 [43, 68, 103, 104]. 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 [73], (ii) folding all sequences belonging to sequence space $\mathcal{I}_\kappa^\ell$ and computation through exhaustive enumeration [37, 38], (iii) computation through statistics of properly chosen samples of sequences [27, 32, 79], and (iv) evaluation through evolutionary optimization [30, 31]. 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 structures from elements [44, 94].[10] The calculation is done by means of the recursion shown in table 2. For large chain lengths ℓ the numbers

[10] 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 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 their high combinatorial probability they amount to less than 10% in the **GC**-case (see also table 3). The loopsize-restriction is much less powerful in reducing the number of shapes than the neglect of structures with isolated base pairs: For chain length $\ell = 30$ we find

TABLE 2 A recursion to calculate the numbers of acceptable RNA secondary structures, $N_S(\ell) = S_\ell^{\left(\min[n_{lp}],\min[n_{st}]\right)}$ [44]. 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 $m+1 \Longrightarrow m$ yields the desired results in the array Ψ_m and uses two auxiliary arrays with the elements Φ_m and Ξ_m which represent the numbers of structures with or without a closing base pair $(1, m)$. One array, e.g. Φ_m, is dispensible but then the formula contains a double sum which is harder to interpret.

Recursion formula:

$$\Xi_{m+1} \;=\; \Psi_m \;+\; \sum_{k=5}^{m-2} \Phi_k \cdot \Psi_{m-k-1}$$

$$\Phi_{m+1} \;=\; \sum_{k=1}^{\lfloor (m-2)/2 \rfloor} \Xi_{m-2k+1}$$

$$\Psi_{m+1} \;=\; \Xi_{m+1} \;+\; \Phi_{m-1}$$

Recursion: $\quad m+1 \;\Longrightarrow\; m$

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: $\quad S_\ell^{(3,2)} \;=\; \Psi_{m=\ell}$

$N_S(\ell)$ are well approximated by the expression:

$$N_S(\ell) \;\approx\; \Sigma(\ell) \;=\; 1.4848 \times \ell^{-3/2} (1.84892)^\ell \;.$$

$\Sigma(\ell)$ 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%, $\ell = 300$ 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^ℓ, and even orders of magnitude smaller than 2^ℓ, the cardinalities of sequence spaces built over four-letter and two-letter alphabets, respectively.

2.15×10^{10} possible secondary structures, 2.41×10^8 structures with loopsize $n_{lp} \geq 3$, and only 760 983 structures with loopsize $n_{lp} \geq 3$ **and** stacksize $n_{st} \geq 2$.

TABLE 3 Comparison of exhaustively folded sequence spaces [35,37,38,76,80]. The values given in parentheses are the counted numbers of actually occurring minimum free energy structures without isolated base pairs which are directly comparable to the total numbers of acceptable structures $N_S(\ell) = S_\ell^{(3,2)}$ (See table 2).

Length	Number of Sequences		Number of Structures			
ℓ	2^ℓ	4^ℓ	$S_\ell^{(3,2)}$	AUGC	GC	AU
7	128	4.29×10^9	2		2	1
10	1 024	1.05×10^6	14		11	1
12	4 096	1.68×10^7	37		31 (29)	1
15	3.28×10^4	1.07×10^9	174		116	2
16	6.55×10^4	4.29×10^9	304	274 (223)	195 (186)	4
17	1.31×10^5	1.73×10^{10}	530		340	8
20	1.05×10^6	1.10×10^{12}	2 741		1 601	35
25	3.36×10^7	1.13×10^{15}	44 695		18 590	164
30	1.07×10^9	1.15×10^{18}	760 983		218 820	1 064

Still we have only an upper limit for the number of shapes actually formed by folding all sequences of a given sequence space $\mathcal{I}_\kappa^\ell$, which evidently obeys $|\mathcal{S}_\kappa(\ell)| \leq N_S(\ell)$. The cardinality of shape space, $|\mathcal{S}_\kappa(\ell)|$, 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 $|\mathcal{S}_{\mathbf{GC}}(30)| = 218\,820$ (table 3).[11] This is only a fraction of 28.8% of all acceptable structures, $N_S(30)$. A comparison of $|\mathcal{S}_{\mathbf{GC}}(30)|$ to the cardinality of sequence space, $|\mathcal{I}_{\mathbf{GC}}^{(30)}| = 1.07 \times 10^9$, shows that the ratio of these numbers is indeed very small, $|\mathcal{S}_{\mathbf{GC}}(30)|/|\mathcal{I}_{\mathbf{GC}}^{(30)}| = 2.045 \times 10^{-4}$. In other words, the mean number of sequences forming the same structure is 4907. In case of four-letter sequences the sequence to structure ratio would be even much larger since we have $|\mathcal{I}_{\mathbf{AUGC}}^{(30)}| = 1.15 \times 10^{18}$ (see also table 4). Thus we are dealing with many more sequences than shapes and, hence, the mapping from sequence space onto shape space is many to one and noninvertible.

[11] This number still contains the structures with isolated base pairs. For $\ell = 16$ we show that these shapes make up less than 10% in the **GC**-case (see table 3).

(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 biassed towards the rare-shape end. Analysis through exhaustive folding [37, 38] yielded a clear result independently of chain lengths ℓ 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 $\ell = 30$ ($\mathbf{GC}_{30}$), 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 $\mathbf{GC}_{30}$ shape space more closely [77]: The most frequent structures are formed by more than 1.5 million sequences which is about 0.15% of sequence space (table 4). 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 ≤ 100 sequences.

(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 [79]. The radius of this shape space covering sphere, $r_{\mathrm{cov}}(\ell)$, can be estimated straightforwardly [76, 77]:

$$r_{\mathrm{cov}}(\ell) = \min\left\{ h = 1, 2, \ldots, \ell \,|\, B_h(\ell, \kappa) \geq \frac{\kappa^\ell}{N_S(\ell)} \right\} \ ,$$

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

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

The covering radius is much smaller than the radius of sequence space ($\ell/2$). For example, it amounts to $r_{\mathrm{cov}} = 15$ for **AUGC**-sequences of chain length $\ell = 100$ and thus on has to search only a fraction of sequence space containing a 4.52×10^{-37}-th of all sequences in order to find all 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

$M_j = \psi^{-1}(S_j) \doteq \{I_k|\psi(I_k) = S_j\}$. A set of sequences can be converted into a graph $\mathcal{G} = (v[\mathcal{G}], e[\mathcal{G}])$ 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 $\mathcal{I}_\kappa^\ell$ (these are the pairs of sequences with Hamming distance $d_{ij}^h = 1$):

$$\begin{aligned}\mathcal{M}_j \;=\; \Big(&v[\mathcal{M}_j] = \big\{I_k \,|\, I_k \in M_j\big\},\\ &e[\mathcal{M}_j] = \big\{(\overline{I_k I_{k'}}) \,|\, I_k, I_{k'} \in M_j \;\text{ and }\; d_{k,k'}^h = 1\big\}\Big).\end{aligned}$$

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 [73, 74]. 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: $\bar{\lambda}_j = \sum_{I_k \in \mathcal{M}_j} \lambda_k/|M_j|$, where λ_k is the number of nearest neighbor sequences of I_k which form shape S_j 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 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:

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

where κ 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 [45].

The predictions on sequence-structure mappings of RNA shapes made by random graph theory were tested through exhaustive folding of entire sequence spaces [35, 37, 38]. 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 $\mathbf{GC}_{30}$ 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 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

TABLE 4 Frequency of common shapes formed by **AUGC**-, **GC**-, and **AU**-sequences of chain length $\ell = 16$ 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.

Stucture	**AUGC**-Alphabet		**GC**-Alphabet		**AU**-Alphabet	
	Rank	Number of Sequences	Rank	Number of Sequences	Rank	Number of Sequences
••••••••••••••••	1	2 709 560 048	9	1427	1	63 488
(((•••)))•••••••	2	52 505 831	13	1301		
•••••••(((•••)))	3	52 376 319	12	1314		
•••••((((•••))))	4	44 544 114	2	2541		
((((•••))))•••••	5	44 273 764	1	2568		
••(((•••)))•••••	6	33 131 192	34	752		
•••••(((•••)))••	7	32 883 686	37	679		
•(((•••)))••••••	8	32 878 614	35	737		
••••••(((•••)))•	9	32 800 711	36	727		
•••(((•••)))••••	10	31 738 681	47	526		
••••(((•••)))•••	11	31 720 954	46	532		
••((((•••))))•••	12	27 886 795	10	1316		
•((((•••))))••••	13	27 835 512	11	1314		
••••((((•••))))•	14	27 791 612	14	1293		
•••((((•••))))••	15	27 778 147	15	1290		
••••((((••••))))	16	26 952 613	3	1895		
((((••••))))••••	17	26 723 146	6	1803		
•••(((((•••)))))	18	24 213 789	5	1880		
(((((•••)))))•••	19	24 047 941	4	1881		
⋮	⋮	⋮	⋮	⋮	⋮	⋮
•((((((•••))))))	65	10 813 722	23	1017	2	1020
((((((•••))))))•	66	10 775 407	24	1015	3	1012
•(((•••••••)))••	67	9 910 874	70	244		
••(((•••••••)))•	68	9 890 910	69	258		
((((((••••))))))	69	8 412 124	25	995	4	16

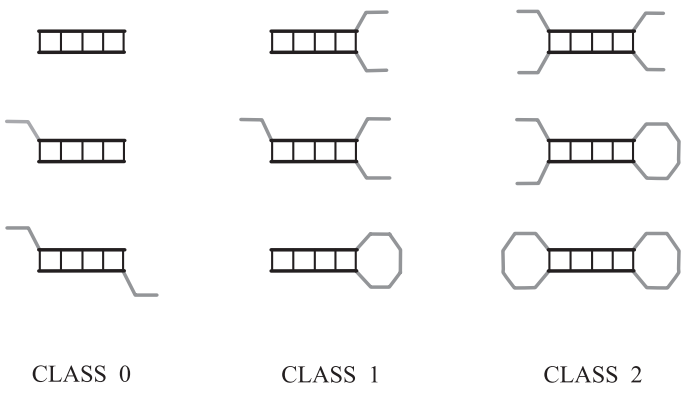

FIGURE 5 **Classes of secondary structures with different distributions in sequence space.** The three classes of structures sketched in the figure differ with respect to the ends of the stacking region. Class **0** structures contain a stack which cannot be extended by closing an additional base pair, class **1** structures can extend the stack on one end, class **2** structures on both ends.

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 membered 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 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 ℓ over an alphabet of size κ in sequence space):

$$B(\ell,h,\kappa) \;=\; \binom{h}{\ell}(\kappa-1)^h \;\Big/\; \kappa^\ell \;,$$

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 arbitarily 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 spontaneously 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 components components 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 occurence 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 componets 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 though merging the two central components.

Finally, we compare the most common shapes formed by **AUGC**-, **GC**- and **AU**-sequences of chain length $\ell = 16$ (table 4) [80]. 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 billions is diffcult 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. ••(((((••••))))), hairpins with internal loops or bulges, e.g. •((•(((•••)))•)) or •(((((•••)))••)), and shapes with two hairpins, e.g. (((•••)))((•••)). In table 4 we listed the most common structures in $\mathcal{I}^{(16)}_{\mathbf{AUGC}}$ and compared them with the shapes formed in the other two alphabets. All structures listed in the table are simple hairpin loops with the only exception being the open chain. The more complex structures are less frequent. The fraction of the open chain structure is a measure of the importance of end effects. In the sequence spaces $\mathcal{I}^{(16)}_{\mathbf{AUGC}}$ and $\mathcal{I}^{(16)}_{\mathbf{AU}}$ the open chain is the most frequent conformation, formed by 63.09% or 96.88% of all sequences, respectively, and hence, end effects dominate. In $\mathcal{I}^{(16)}_{\mathbf{GC}}$ the open chain has rank no. 9 and is formed by only 2.18%. 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 three structures of $\mathcal{S}_{\mathbf{AU}}(16)$ are tri- and tetraloops attached to a stack of six base pairs. Interestingly, they are less frequent in the other two sequence spaces, ranks no. 23, no. 24, and no. 25 in $\mathcal{I}^{(16)}_{\mathbf{GC}}$, and ranks no. 65, no. 66, and no. 69 in $\mathcal{I}^{(16)}_{\mathbf{AUGC}}$, respectively. Comparing $\mathcal{I}^{(16)}_{\mathbf{AUGC}}$ and $\mathcal{I}^{(16)}_{\mathbf{GC}}$ we

see that the **AUGC**-alphabet sustains about 40% more shapes than the **GC**-alphabet, 274 versus 195 (table 3). Although the ranking of shapes shows substantial differences, forteen out of the first twenty shapes coincide. 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.

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. The chemical reaction mechanism contains replication and mutation steps apart from flow terms. 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 [33,34]. 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 [28, 29] 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 never observed again in subsequent identical simulations,[12] and (iii) sharp 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 [45]. 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

[12] By "two simulation experiments under identical conditions" we mean that everything was kept constant except the seeds for the random number generators.

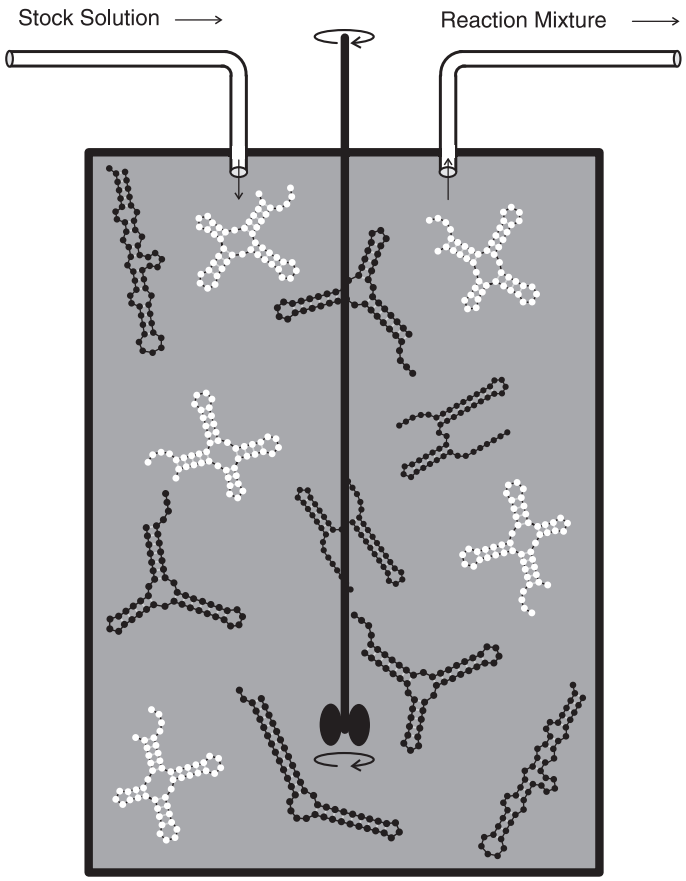

FIGURE 6 **The flow reactor as a device for RNA structure optimization.** RNA molecules with different shapes are produced through replication and mutation. New sequences obtained by mutation are folded into minimum free energy secondary structures. Replication rate constants are computed from structures by means of predefined rules (see text). For example, the replication rate is a function of the distance to a target stucture 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 ℓ of the RNA molecules as well as the mutation rate p.

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 $a(S_j; S_\tau) = (\alpha + d^s_{j\tau}/\ell)^{-1}$ was used, where α is some constant, ℓ the chain length of the RNA, and $d^s_{j\tau}$ the distance between structure S_j and the target structure S_τ [31]. Many measures of

distance between structures are conceivable [27], a particularly simple one is the Hamming distance between the short-hand "parentheses notations" (Sect. 3) of the two shapes. 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 constants, fitness functions, and distance measures.

Optimization of RNA structures was studied through simulations of the evolution of a population in the flow reactor [30, 31]. 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,

$$\overline{d_\tau^s}(t) \;=\; \sum_{j=1}^{n_s} p_j^s(t)\, d_{j\tau}^s \;, \tag{10}$$

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, $p_j^s(t) = N_j^s(t)/N$ is the frequency of structure S_j, and $N_j^s(t)$ 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 intial 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 uninterrupred 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 a series of shapes and times of first appearence which ultimately ends in the initial population.[13] The full relay series of the computer experiment of figure 7 is shown in figure 8. It contains 42 shapes produced through $n_{\mathrm{rl}} = 41$ 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 occurence 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

[13] 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 appearences. (ii) A relay series is not a genealogy which is the full recording of a line of genotypes in parent-offspring relation.

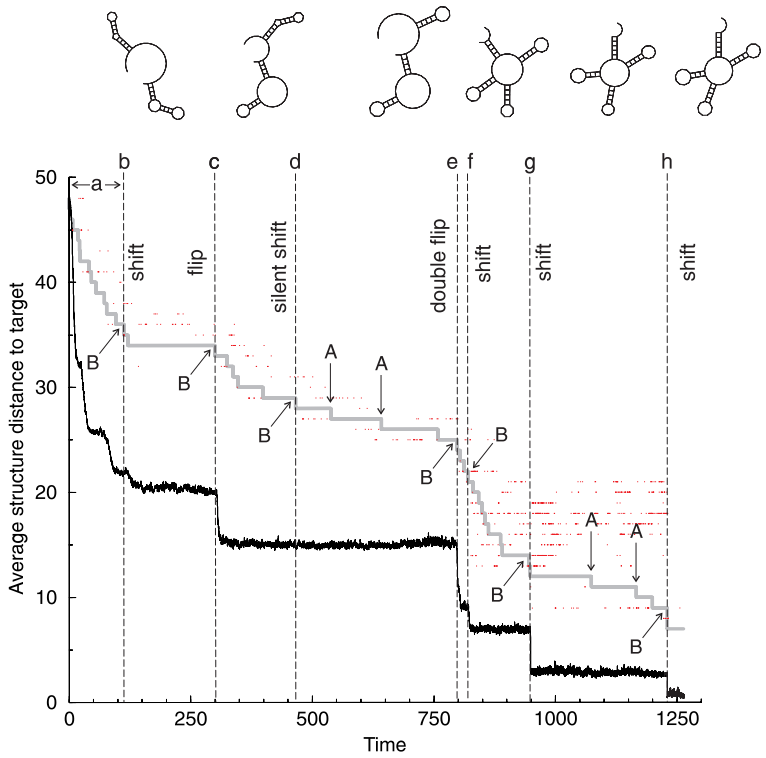

FIGURE 7 **The recording of an RNA structure optimization experiment in the flow reactor.** The computer experiment starts from a homogeneous initial population of about 1000 RNA molecules with an arbitrarily chosen structure and leads to a quasispecies like distribution around the target shape. Fitness expressed as replication rate is computed as a function of the distance between current (S_j) and target structure (S_τ), $d^s_{j\tau}$ (For details see text). The target structure was chosen to be the clover leaf of $\mathrm{tRNA}^{\mathrm{phe}}$ ($\ell = 76$). The mean distance to the target structure of the entire population, $\overline{d^s_\tau}(t)$ 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 1.1×10^7 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**). An more or less well defined intial period of about one hundred time units is characterized by fast decrease in the distance to the target (**a**). The individual discontinous 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.

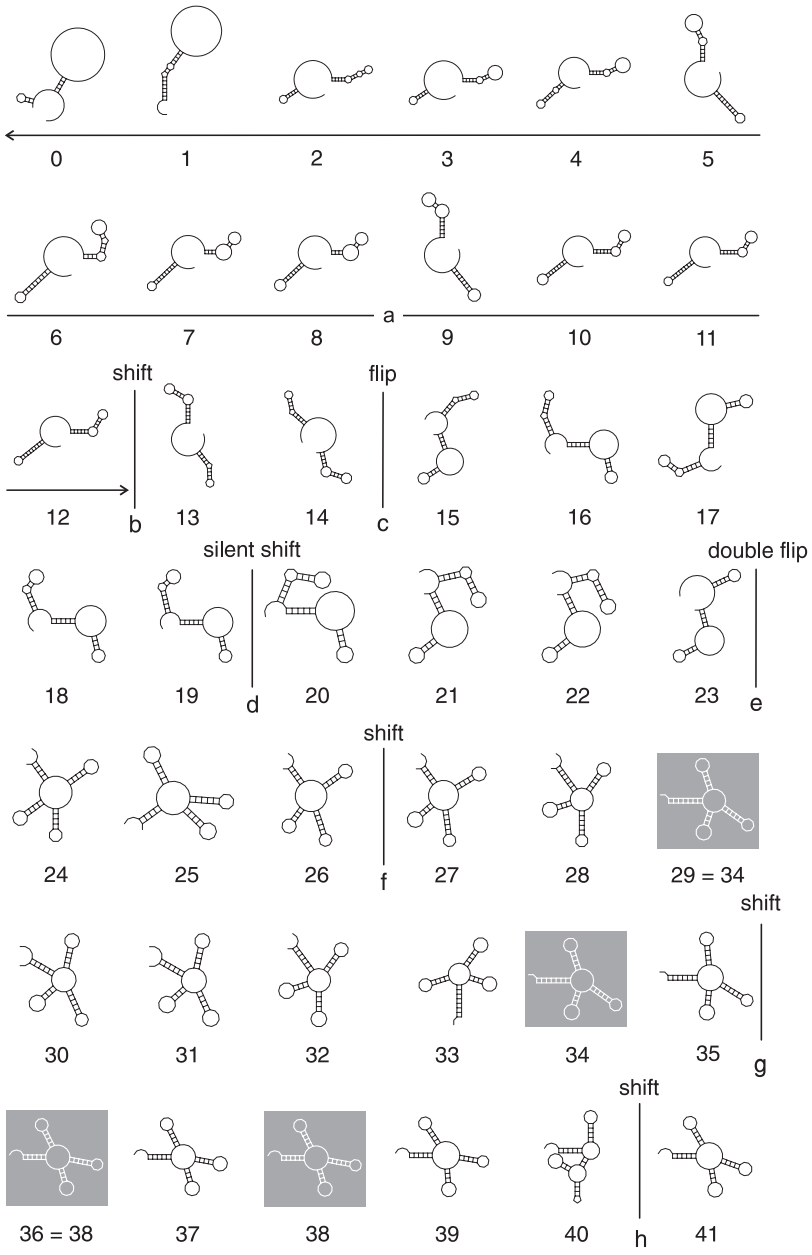

FIGURE 8 **The relay series of the *in silico* optimization experiment described in figure 7**. For details see text. It is worth noticing that a given shape may appear twice or more often in a relay series. Examples are the shapes $29 \equiv 34$ and $36 \equiv 38$.

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 [96]. 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: $\overline{d^s_{\mathrm{in},\tau}} = \sum_{j=1}^{N(t_{\mathrm{in}})} d^s_{j,\tau}$. Herein $d^s_{j,\tau}$ is the structure distance between $S_j = \psi(I_j)$ and the target shape S_τ, and t_{in} the time at which the intital phase ends.

Population Size N	Number of Runs	Number of Relay Steps n_{rl}	Number of Transitions n_{mt}	Initial Phase $\overline{d^s_{\mathrm{in},\tau}}$
1 000	10	120.1 ± 114.0	6.5 ± 1.7	17.6 ± 2.3
2 000	13	66.3 ± 25.8	6.5 ± 1.7	18.5 ± 2.3
3 000	12	41.9 ± 16.6	6.3 ± 2.2	17.6 ± 2.4
10 000	17	37.8 ± 11.8	5.7 ± 1.3	16.5 ± 1.0

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 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 [31]). 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 intial shape.

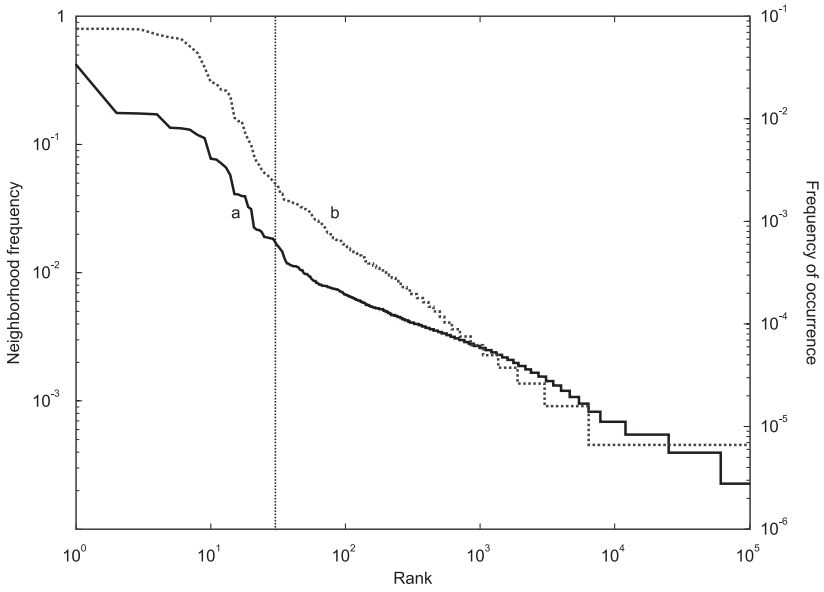

FIGURE 9 **Statistics of shapes in the boundary of tRNA$^{\mathrm{phe}}$.** The basis of the statistics are 2199 sequences folding into the clover-leaf structure shown in figure 2. All their one-error mutants, 501 372 in number, were folded. A fraction of 28% formed the same clover-leaf as the reference sequence and thus belonged to the neutral network. The remaining 358 525 seqeunces folded into 141 907 distinct shapes. Curve **a** is a log-log plot of the rank ordered frequency of occurence (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 occurence (left) is distiguished from the power-law distribution (right), which is typical for scaling according to Zipf's law [100].

(ii) **Discontinuous transitions** (**B**). They involve major stuctural changes leading to globally rare and only locally frequent structures. Accordingly, discontinuous transitions require special sequences that allow major structural changes to occur on single point mutations.

Simulations show an initial period ($0 \leq t \leq t_{\mathrm{in}}$; marked **a** in figure 7) of cascading discontinuous and continuous transitions followed by a stepwise optimization process with apparent regularities. Each epoch or quasi-stationary phase of evolution ends with a discontinuous transition. Discontinuous transitions (**b** to **h**), however, occur only rarely within quiescent periods.[14] Every discontinuous transition is followed by a cascade of continuous transitions

[14]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.

which are accompanied by fitness increase. Then, the population approaches the next plateau corresponding to an epoch of neutral evolution at approximately 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 discontinuous 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 neutral 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 of evolutionary nearness between phenotypes which is based on the concept of neutral networks [31]. 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, $B_j = \mathrm{bd}(M_j)$, 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, Σ_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.[15] 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 [100], which implies that the log(frequency)/ log(rank)-plot is a straight line. The results 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. 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

[15] At higher error-rates it might useful to define two-error, three-error or, in general n-error boundaries [31].

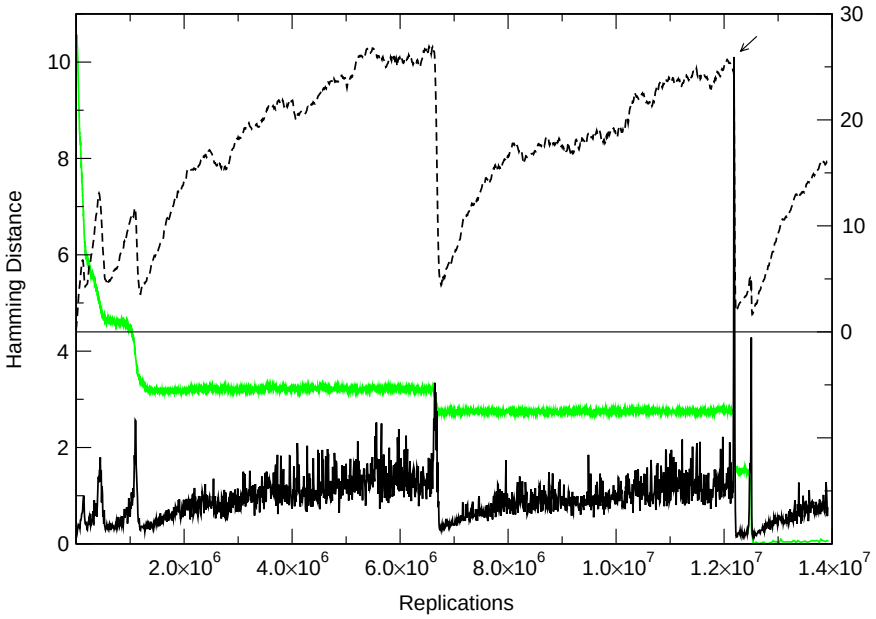

FIGURE 10 **Variability in genotype space during punctuated evolution**. Shown are the results of a simulation of RNA optimization towards a tRNA target (analogous to the run in figure 7) with population size $n = 3000$ and mutation rate $p = 0.001$ per site and replication. The figure contains two plots with different measures of genetic diversity, $d_P(t, \Delta t)$ and $d_C(t, \Delta t)$ with $\Delta t = 8\,000$ replications, against time, which is expressend 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 $d_C(t, 8\,000)$ 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 incerases 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 8 000 replications.

$\mathcal{M}_j$ (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 (sta-

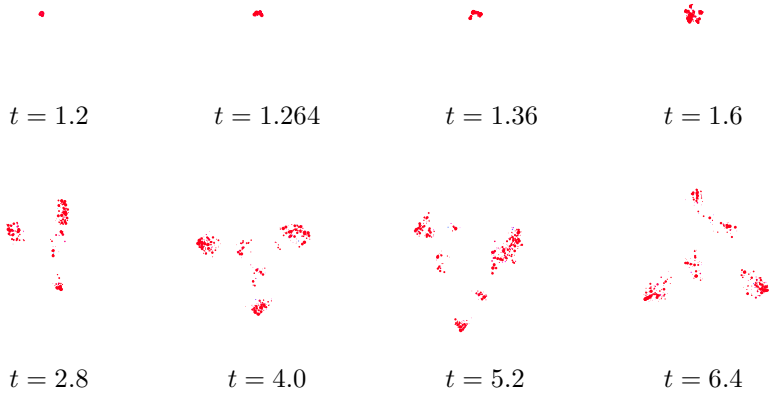

FIGURE 11 **Spreading of a population in genotype space during a quasi-stationary epoch**. The individual figures are snapshots of the genotype distribution at times corresponding to 1.2, 1.264, 1.36, 1.6, 2.8, 4.0, 5.2, and 6.4×10^6 replications. In order to visualize spreading genotype distributions were transformed to principal axes and individual sequences were projected onto the planned 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 clusters which diffuse radially away from the center of the distribution (See also the model on neutral evolution discussed in [12, 45]).

tistically) **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 ε 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 fulfil 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 fulfil 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 two networks of different size in sequence space are considered. The larger network can occupy a fairly high percentage of the positions in the boundary of the smaller network whereas, at the same time, the smaller one is present only at low

frequency in the boundary of the larger network.[16] The proper mathematical context for accessibility of phenotypes as well as continuity and discontinuity in evolution is still being developed [9, 10] 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 $t + \Delta t$,

$$d_C(t, \Delta t) \;=\; \sum_{k=1}^{\ell} \sqrt{\sum_{j=\mathbf{A},\mathbf{U},\mathbf{G},\mathbf{C}} \left(\pi_j^{(k)}(t) \;-\; \pi_j^{(k)}(t+\Delta t)\right)^2 \Big/\; 2}\;.$$

The vector $\vec{\pi}^{(k)}(t) = \{\pi_{\mathbf{A}}^{(k)}, \pi_{\mathbf{U}}^{(k)}, \pi_{\mathbf{G}}^{(k)}, \pi_{\mathbf{C}}^{(k)}\}$ is the square-normalized distribution of nucleotides at position k: $\pi_i^{(k)} = \alpha_i^{(k)} \;/\; \sqrt{\sum_{j=\mathbf{A},\mathbf{U},\mathbf{G},\mathbf{C}} \left(\alpha_j^{(k)}\right)^2}$ with $\sum_{j=\mathbf{A},\mathbf{U},\mathbf{G},\mathbf{C}} \alpha_j^{(k)} = 1$. 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.[17] The size of mutant cloud increases with time

[16] For examples of pairs neutral networks of RNA molecules lacking symmetry in the statistical neighborhood see [30]. We mention here only one case: a clover-leaf tRNA shape (S_{tRNA}) and a conformation with three hairpins which originates from the clover-leaf through opening of the terminal stack ($S_{3\mathrm{hp}}$; the terminal stack is the upper vertical stack of the secondary structure in figure 2). The latter, $S_{3\mathrm{hp}}$, is near S_{tRNA} in a statistical sence, but the inverse is not true, S_{tRNA} is not statistically near $S_{3\mathrm{hp}}$. The interpretation of this asymmetry is straighforward. Let $\gamma(S_{\mathrm{tRNA}}, S_{3\mathrm{hp}})$ be the number of Hamming distance one contacts between the two networks. The sizes of the two networks are very different: $|M_{\mathrm{tRNA}}| \ll |M_{3\mathrm{hp}}|$, and hence, the boundaries too, $|B_{\mathrm{tRNA}}| \ll |B_{3\mathrm{hp}}|$, which leads to

$$\rho(S_{\mathrm{tRNA}}; S_{3\mathrm{hp}}) = \gamma(S_{\mathrm{tRNA}}, S_{3\mathrm{hp}})/|B_{3\mathrm{hp}}| \gg \rho(S_{3\mathrm{hp}}; S_{\mathrm{tRNA}}) = \gamma(S_{3\mathrm{hp}}, S_{\mathrm{tRNA}})/|B_{\mathrm{tRNA}}|$$

and thus the statistical neighborhood relation does not commute, q.e.d.

[17] The delay time Δt was chosen to be identical in both quantities, $d_P(t, \Delta t)$ and $d_C(t, \Delta t)$. We remark that $d_P(t, \Delta t)$ chances hardly within the time span considered: $d_P(t, 8\,000)$ is almost identical with $d_P(t, 0)$.

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 8 000 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 and 6.4×10^6 replications (figure 11). After having passed the bottleneck of the previous discontinuous transition the population starts instantaneously to expand. At a time corresponding to 2.8×10^6 replications the population breaks up into subpopulations which diverge further in sequence space. This finding is in agreement with previous results [12,45] 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 [70], 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 variation 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, $\overline{d^s_{\mathrm{in},\tau}}$, 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 [78]. 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 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 [101] 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 [31]. 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 [70].
(iv) The landscape concept, originally introduced by Sewall Wright [98] 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 [5, 20, 21, 23]. 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 know to be suitable for this purpose are the nucleic acids. No wonder that evolution of molecules and its applications to biotechnology [25, 36, 87, 95] were more or less restricted to the use of RNA and DNA. Design of proteins or organic molecules by means of selection techniques requires always direct intervention by the experimentalists.

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 alphabet, although it is straightforward to show that such molecules do exist and are stable (See, for example [31], 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 random drift" builts up the properties that are not encoded in the conformation of the ground state, and thus diffusion on networks (which are only neutral with respect to the mfe-structure) may also generate genetic information.

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 Gttingen, 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

[1] Alves, D. and J. F. Fontanari, "A Population Genetic Approach to the Quasispecies Model". *Phys. Rev. E* **54** (1996):4048–4053.

[2] Alves, D. and J. F. Fontanari, "Error Thresholds in Finite Populations". *Phys. Rev. E* **57** (1998):7008–7013.

[3] Batey, R. T., R. P. Rambo, and J. A. Doudna, "Tertiary Motifs in Structure and Folding of RNA". *Angew. Chem. Int. Ed.* **38** (1999):2326–2343.

[4] Biebricher, C. K. and W. C. Gardiner, "Molecular Evolution of RNA *in vitro*". *Biophys. Chem.* **66** (1997):179–192.

[5] Brakmann, S., "On the Generation of Information as Motive Power for Molecular Evolution". *Biophys. Chem.* **66** (1997):133–143.

[6] Breton, N., C. Jacob, and P. Daegelen, "Prediction of Sequentially Optimal RNA Secondary Structures". *J. Biomol. Struct. Dynam.* **14** (1997):727–740.

[7] Campos, P. R. A. and J. F. Fontanari, "Finite-size Scaling of the Quasispecies Model". *Phys. Rev. E* **58** (1998):2664–2667.

[8] Campos, P. R. A. and J. F. Fontanari, "Finite-size Scaling of the Quasispecies Model". *J. Phys. A: Math. Gen.* **32** (1999):L1–L7.

[9] Cupal, J., S. Kopp, and P. F. Stadler, "RNA Shape Space Topology". 1999, los Alamos National Laboratory, Preprint LA-UR 99:1324.

[10] Cupal, J., P. Schuster, and P. F. Stadler, "Topology in Phenoptype Space". In *Computer Science in Biology*, 9–15, GCB'99 Proceedings, Hannover, DE: Univ. Bielefeld, 1999.

[11] Demetrius, L., P. Schuster, and K. Sigmund, "Polynucleotide Evolution and Branching Processes". *Bull. Math. Biol.* **47** (1985):239–262.
[12] Derrida, B. and L. Peliti, "Evolution in a Flat Fitness Landscape". *Bull. Math. Biol.* **53** (1991):355–382.
[13] Domingo, E., "Biological Significance of Viral Quasispecies". *Viral Hepatitis Rev* **2** (1996):247–261.
[14] Domingo, E. and J. J. Holland, "RNA Virus Mutations and Fitness for Survival". *Ann. Rev. Microbiol* **51** (1997):151–178.
[15] Domingo, E., L. Menéndez-Arias, M. E. Quinoñes-Mateu, A. Holguín, M. Gutierrez-Rivas, M. A. Martínez, J. Quer, and J. J. Holland, "Viral Quasispecies and the Problem of Vaccine-escape and Drug-resistance mutants". *Prog. Drug. Res* **48** (1997):99–128.
[16] Drake, J. W., "A Constant Rate of Spontaneous Mutation in DNA-based Microbes". *Proc. Natl. Acad. Sci. USA* **88** (1991):7160–7164.
[17] Drake, J. W., "Rates of Spontaneous Mutation among RNA Viruses". *Proc. Natl. Acad. Sci. USA* **90** (1993):4171–4175.
[18] Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow, "Rates of Spontaneous Mutation". *Genetics* **148** (1998):1667–1686.
[19] Ebeling, W. and R. Mahnke, "Kinetics of Molecular Replication and Selection". *Problems of Contemporary Biophysics (Zagadnienia Biofizyki Współczesnej)* **4** (1979):119–128.
[20] Eigen, M., "Selforganization of Matter and the Evolution of Biological Macromolecules". *Naturwissenschaften* **58** (1971):465–523.
[21] Eigen, M., "The Origin of Genetic Information: Viruses as Models". *Gene* **135** (1993):37–47.
[22] Eigen, M., J. McCaskill, and P. Schuster, "The Molecular Quasispecies." *Adv. Chem. Phys.* **75** (1989):149–263.
[23] Eigen, M. and P. Schuster, "The Hypercycle. A Principle of Natural Self-Organization. Part A: Emergence of the Hypercycle". *Naturwissenschaften* **64** (1977):541–565.
[24] Elena, S. F., V. S. Cooper, and R. E. Lenski, "Punctuated Evolution Caused by Selection of Rare Beneficial Mutants". *Science* **272** (1996):1802–1804.
[25] Ellington, A. D., "RNA selection. Aptamers achieve the desired recognition." *Curr. Biol.* **4** (1994):427–429.
[26] Flamm, C., W. Fontana, I. L. Hofacker, and P. Schuster, "Elementary Step Dynamics of RNA Folding". *RNA* **6** (2000), in press.
[27] Fontana, W., D. A. M. Konings, P. F. Stadler, and P. Schuster, "Statistics of RNA secondary structures." *Biopolymers* **33** (1993):1389–1404.
[28] Fontana, W., W. Schnabl, and P. Schuster, "Physical Aspects of Evolutionary Optimization and Adaptation." *Phys. Rev. A* **40** (1989):3301–3321.

[29] Fontana, W. and P. Schuster, "A Computer Model of Evolutionary Optimization". *Biophys. Chem.* **26** (1987):123–147.

[30] Fontana, W. and P. Schuster, "Continuity in Evolution. On the Nature of Transitions". *Science* **280** (1998):1451–1455.

[31] Fontana, W. and P. Schuster, "Shaping Space. The Possible and the Attainable in RNA Genotype-Phenotype Mapping". *J. Theor. Biol.* **194** (1998):491–515.

[32] Fontana, W., P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, and P. Schuster, "RNA Folding and Combinatory Landscapes." *Phys. Rev. E* **47** (1993):2083–2099.

[33] Gillespie, D. T., "A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions". *J. Comp. Phys.* **22** (1976):403–434.

[34] Gillespie, D. T., "Exact Stochastic Simulation of Coupled Chemical Reactions". *J. Phys. Chem.* **81** (1977):2340–2361.

[35] Göbel, U., S. Kopp, and P. Schuster, "Complete Sequence-Secondary Structure Mapping of Oligo-Ribonucleotides of Chain Length $n = 16$". 1999, preprint.

[36] Gold, L., C. Tuerk, P. Allen, J. Binkley, D. Brown, L. Green, S. MacDougal, D. Schneider, D. Tasset, and S. R. Eddy, "RNA: The Shape of Things to Come". In *The RNA World*, edited by R. F. Gesteland and J. F. Atkins, 497–509, Plainview, NY: Cold Spring Harbor Laboratory Press, 1993.

[37] Grüner, W., R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, and P. Schuster, "Analysis of RNA Sequence Structure Maps by Exhaustive Enumeration.
II. Structures of Neutral Networks and Shape Space Covering". *Mh.Chemie* **127** (1996):375–389.

[38] Grüner, W., R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, and P. Schuster, "Analysis of RNA Sequence Structure Maps by Exhaustive Enumeration.
I. Neutral Networks". *Mh.Chemie* **127** (1996):355–374.

[39] Gultyaev, A. P., F. H. D. van Batenburg, and C. W. A. Pleij, "The Computer Simulation of RNA Folding Pathways using a Genetic Algorithm". *J.Mol.Biol.* **250** (1995):37–51.

[40] Gultyaev, A. P., F. H. D. van Batenburg, and C. W. A. Pleij, "Dynamic Competition between Alternative Structures in Viroid RNAs Simulated by an RNA Folding Algorithm". *J.Mol.Biol.* **276** (1998):43–55.

[41] Hamming, R. W., "Error Detecting and Error Correcting Codes". *Bell Syst. Tech. J.* **29** (1950):147–160.

[42] Hamming, R. W., *Coding and Information Theory*. Englewood Cliffs, NJ: Prentice Hall, 2nd edition, 1989.

[43] Hofacker, I. L., W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, "Fast Folding and Comparison of RNA Secondary Structures." *Mh. Chem.* **125** (1994):167–188.
[44] Hofacker, I. L., P. Schuster, and P. F. Stadler, "Combinatorics of RNA Secondary Structures". *Discr. Appl. Math.* **88** (1998):207–237.
[45] Huynen, M. A., P. F. Stadler, and W. Fontana, "Smoothness within Ruggedness: The role of Neutrality in Adaptation". *Proc. Natl. Acad. Sci. USA* **93** (1996):397–401.
[46] Jacob, C., N. Breton, and P. Daegelen, "Stochastic Theories of the Activated Complex and the Activated Collision: The RNA Example". *J. Chem. Phys.* **107** (1997):2903–2912.
[47] Jacob, C., N. Breton, P. Daegelen, and J. Peccoud, "Probability Distribution of the Chemical States of a Closed System and Thermodynamic Law of Mass Action form Kinetics: The RNA Example". *J. Chem. Phys.* **107** (1997):2913–2919.
[48] Jones, B. L., R. H. Enns, and S. S. Rangnekar, "On the Theory of Selection of Coupled Macromolecular Systems". *Bull. Math. Biol.* **38** (1976):15–28.
[49] Jones, B. L. and H. K. Leung, "Stochastic Analysis of a Nonlinear Model for Selection of Biological Macromolecules". *Bull. Math. Biol.* **43** (1981):665–680.
[50] Kauffman, S. A., *The Origins of Order. Self-Organization and Selection in Evolution*. Oxford, UK: Oxford University Press, 1993.
[51] Kimura, M., "Evolutionary Rate at the Molecular Level". *Nature* **217** (1968):624–626.
[52] Kimura, M., *The Neutral Theory of Molecular Evolution*. Cambridge, UK: Cambridge University Press, 1983.
[53] Lenski, R. E. and M. Travisano, "Dynamics of Adaptation and Diversification: A 10 000-generation Experiment with Bacterial Populations". *Proc. Natl. Acad. Sci. USA* **91** (1994):6808–6814.
[54] Leung, H. K., "Stability Analysis of a Stochastic Model for Biomolecular Selection". *Bull. Math. Biol.* **46** (1984):399–406.
[55] Leung, H. K., "Expansion of the Master Equation for a Biomolecular Selection Model". *Bull. Math. Biol.* **47** (1985):231–238.
[56] Leuthäusser, I., "Statistical Mechanics of Eigen's Evolution Model". *J. Stat. Phys.* **48** (1987):343–360.
[57] Martinez, H. M., "An RNA Folding Rule". *Nucl. Acids Res.* **12** (1984):323–334.
[58] McCaskill, J. S., "A Localization Threshold for Macromolecular Quasi-Species from Continuously Distributed Replication Rates". *J. Chem. Phys.* **80** (1984):5194–5202.
[59] McCaskill, J. S., "The Equilibrium Partition Function and Base Pair Binding Probabilities for RNA Secondary Structures". *Biopolymers* **29** (1990):1105–1119.

[60] Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of State Calculations by Fast Computing Machines". *J. Chem. Phys.* **21** (1953):1087–1092.

[61] Mills, D. R., R. L. Peterson, and S. Spiegelman, "An Extracellular Darwinian Experiment With a Self-duplicating Nucleic Acid Molecule". *Proc. Natl. Acad. Sci. USA* **58** (1967):217–224.

[62] Mironov, A. and A. Kister, "A Kinetic Approach to the Prediction of RNA Secondary Structures". *J. Biomol. Struct. Dyn.* **2** (1985):953–962.

[63] Mironov, A. and A. Kister, "RNA Secondary Structure Formation during Transcription". *J. Biomol. Struct. Dyn.* **4** (1985):1–9.

[64] Mironov, A. and V. F. Lebedev, "A Kinetic Model of RNA Folding". *BioSystems* **30** (1993):49–56.

[65] Moran, P. A. P., "The effect of selection in haploid genetic populations". *Proc. Camb. Phil. Soc.* **54** (1958):463–474.

[66] Morgan, S. R. and P. G. Higgs, "Evidence for Kinetic Effects in the Folding of Large RNA Molecules". *J. Chem. Phys.* **105** (1996):7152–7157.

[67] Nowak, M. and P. Schuster, "Error Thresholds of Replication in Finite Populations. Mutation Frequencies and the Onset of Muller's Ratchet." *J. Theor. Biol.* **137** (1989):375–395.

[68] Nussinov, R. and A. B. Jacobson, "Fast Algorithm for Predicting the Secondary Structure of Single-stranded RNA". *Proc. Natl. Acad. Sci. USA* **77** (1980):6309–6313.

[69] Ohta, T., "The Nearly Neutral Theory of Molecular Evolution". *Annu. Rev. Ecol. Syst.* **23** (1992):263–286.

[70] Papadopoulos, D., D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski, and M. Blot, "Genomic Evolution during a 10 000-generation Experiment with Bacteria". *Proc. Natl. Acad. Sci. USA* **96** (1999):3807–3812.

[71] Pütz, J., J. D. Puglisi, C. Florentz, and R. Giegé, "Identity Elements for Specific Aminoacylation of Yeast $tRNA^{asp}$ by Cognate Aspartyl tRNA Synthetase". *Science* **252** (1991):1696–1699.

[72] Reidys, C., C. Forst, and P. Schuster, "Replication and Mutation on Neutral Networks". *J. Math. Biol.* (2000), in press.

[73] Reidys, C., P. F. Stadler, and P. Schuster, "Generic Properties of Combinatory Maps. Neutral Networks of RNA Secondary Structure." *Bull. Math. Biol.* **59** (1997):339–397.

[74] Reidys, C. M., "Random Induced Subgraphs of Generalized n-Cubes". *Adv. Appl. Math.* **19** (1997):360–377.

[75] Rohde, N., H. Daum, and C. K. Biebricher, "The mutant distribution of an RNA species replicated by $Q\beta$ replicase." *J. Mol. Biol.* **249** (1995):754–762.

[76] Schuster, P., "How to Search for RNA Structures. Theoretical Concepts in Evolutionary Biotechnology." *J. Biotechnol.* **41** (1995):239–257.

[77] Schuster, P., "Landscapes and Molecular Evolution". *Physica D* **107** (1997):351–365.
[78] Schuster, P. and W. Fontana, "Chance and Necessity in Evolution: Lessons from RNA". *Physica D* **133** (1999):427–452.
[79] Schuster, P., W. Fontana, P. F. Stadler, and I. L. Hofacker, "From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures". *Proc. Roy.Ṡoc. Lond. B* **255** (1994):279–284.
[80] Schuster, P. and P. F. Stadler, "Discrete Models of Biopolymers". In *Handbook of Computational Chemistry*, edited by M. J. C. Crabbe, M. Drew, and A. Konopka, New York: Marcel Dekker, 2000, in press.
[81] Schuster, P. and J. Swetina, "Stationary Mutant Distribution and Evolutionary Optimization". *Bull. Math. Biol.* **50** (1988):635–660.
[82] Spiegelman, S., "An Approach to the Experimental Analysis of Precellular Evolution". *Quart. Rev. Biophys.* **4** (1971):213–253.
[83] Stadler, P. F., "Fitness Landscapes Arising from the Sequence-Structure Maps of Biopolymers". *J. Mol. Struct. (Theochem)* **463** (1999):7–19.
[84] Suvernev, A. A. and P. A. Frantsuzov, "Statistical Description of Nucleic Acid Secondary Structure Folding". *J. Biomol. Struct. Dynam.* **13** (1995):135–144.
[85] Swetina, J., "First and Second Moments and the Mean Hamming Distance in a Stochastic Replication-Mutation Model for Biological Macromolecules". *J. Math. Biol.* **27** (1989):463–483.
[86] Swetina, J. and P. Schuster, "Self-Replication with Errors - A Model for Polynucleotide Replication". *Biophys. Chem.* **16** (1982):329–345.
[87] Szostak, J. W. and A. D. Ellington, "*In vitro* Selection of Functional RNA Sequences". In *The RNA World*, edited by R. F. Gesteland and J. F. Atkins, 511–533, Plainview, NY: Cold Spring Harbor Laboratory Press, 1993.
[88] Tacker, M., W. Fontana, P. F. Stadler, and P. Schuster, "Statistics of RNA Melting Kinetics." *Eur.Biophys.J.* **23** (1994):29–38.
[89] Tarazona, P., "Error Threshold for Molecular Quasispecies as Phase Transitions: From Simple Landscapes to Spin-Glass Models". *Phys. Rev. A* **45** (1992):6038–6050.
[90] Thompson, C. J. and J. L. McBride, "On Eigen's Theory of the Self-Organization of Matter and the Evolution of Biological Macromolecules". *Math. Biosci.* **21** (1974):127–142.
[91] van Kampen, N. G., "The Expansion of the Master Equation". *Adv. Chem. Phys.* **34** (1976):245–309.
[92] van Nimwegen, E., *The Statistical Dynamics of Epochal Evolution*. Ph.D. thesis, Universiteit Utrecht, Utrecht, NL, 1999.
[93] van Nimwegen, E., J. P. Crutchfield, and M. Mitchell, "Finite Populations Induce Metastability in Evolutionary Search". *Phys. Lett. A* **229** (1997):144–150.

[94] Waterman, M. S., "Secondary Structure of Single-Stranded Nucleic Acids". *Adv. Math. Suppl. Studies* **1** (1978):167–212.

[95] Watts, A. and G. Schwarz, editors, *Evolutionary Biotechnology – From Theory to Experiment*, vol. 66/2-3 of *Biophysical Chemistry*. Amsterdam: Elesvier, 1997.

[96] Wernitznig, A., C. Flamm, and P. Schuster, "RNA Evolution *in silico*: The Flow Reactor". 2000, preprint.

[97] Wiehe, T., E. Baake, and P. Schuster, "Error Propagation in Reproduction of Diploid Organisms. A Case Study in Single Peaked Landscapes." *J. Theor. Biol.* **177** (1995):1–15.

[98] Wright, S., "The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution". In *Int. Proceedings of the Sixth International Congress on Genetics*, edited by D. F. Jones, vol. 1, 356–366, 1932.

[99] Wuchty, S., W. Fontana, I. L. Hofacker, and P. Schuster, "Complete Suboptimal Folding of RNA and the Stability of Secondary Structures". *Biopolymers* **49** (1999):145–165.

[100] Zipf, G., *Human Behaviour and the Principle of Least Effort*. Reading, MA: Addison-Wesley, 1949.

[101] Zuckerkandl, E., "Neutral and Nonneutral Mutations: The Creative Mix – Evolution of Complexity in Gene Interaction Systems". *J. Mol. Evol.* **44 (Suppl.1)** (1997):S2–S8.

[102] Zuker, M., "On Finding All Suboptimal Foldings of an RNA Molecule". *Science* **244** (1989):48–52.

[103] Zuker, M. and D. Sankoff, "RNA secondary structures and their prediction". *Bull. Math. Biol.* **46** (1984):591–621.

[104] Zuker, M. and P. Stiegler, "Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information". *Nucleic Acids Research* **9** (1981):133–148.