EVOLUTION AT MOLECULAR RESOLUTION[a]

P. SCHUSTER
Institute for Theoretical Chemistry, University of Vienna,
Währingerstraße 17, A-1090 Wien, Austria
E-mail: pks@tbi.univie.ac.at

Conventional population genetics is extended by support dynamics and genotype-phenotype mapping in order to conceive a comprehensive dynamical model of evolution. Support dynamics describes migration of populations through genotype space. The relation between genotypes and phenotypes is a core issue of evolution. In the simplest conceivable case, *in vitro* evolution of RNA molecules, both phenomena can be incorporated into computer simulations. Application of replication-mutation kinetics to processes in the space of genotypes led to the notion of **quasispecies** which has been applied successfully to evolution of molecules and viruses. In molecular evolution mapping of genotypes into phenotypes is tantamount to sequence-structure relations of RNA molecules. Systematic studies were performed on secondary structures. They revealed a number of regularities which are reported. The number of sequences is much larger than the number of secondary structures and thus neutrality is a central issue of sequence-structure mappings. Evolution of populations of RNA molecules towards a predefined target structure were carried out and analyzed in molecular detail. The results derived for RNA molecules suggested to define a statistical relation of nearness between phenotypes which constitutes a kind of **statistical topology**. This probabilistic concept of neighborhood in sequence space can be generalized and appears to be of widespread validity in evolution.

1 Introduction

Biological evolution is too complex and too slow for experimental investigation. In order to make evolutionary phenomena accessible to systematic studies one needs (i) to reduce generation times in order to speed up evolution, (ii) to minimize complexity of phenotypes in order to allow for an analysis of genotype-phenotype relations, and (iii) to shorten genotype lengths in order to keep possible diversity below a certain limit. All three conditions are fulfilled, for example, by test-tube experiments on optimization of RNA molecules.[1] Evolution of molecules in the test tube is indeed the simplest and the only currently known realistic system that allows to study the mechanisms of biological evolution at molecular resolution. Both, the experimental approach and the development of theory, have reached a point from where on systematic studies and global investigations of the rules underlying the dynamics of evolutionary processes are required in order to make progress in the understanding

[a] Published in: Leif Matsson, ed. *Nolinear Cooperative Phenomena in Biological Systems*, pp. 86-112. World Scientific, Singapore 1998.

of the phenomenon and the design of new conclusive experiments. Although many successful studies have already been reported and it is generally accepted now that evolutionary optimization of molecular properties and functions does not require cellular life, the design of efficient experiments leading to optimal molecules is anything but trivial. In addition, the currently available data on the evolution of primitive systems call for a comprehensive theoretical frame that allows to put them into proper context.

2 Evolution of molecules

The term molecular evolution is currently used for two related but nevertheless distinct fields of research: (i) The fast increasing availability of sequences of natural biomolecules allows to compare sequences of biomolecules with the same function in different organisms and to reconstruct phylogenetic trees from these molecular data.[2] The theoretical frame of this approach was provided by the neutral theory of evolution.[3,4] (ii) Molecular evolution can also be understood as "evolution of molecules" in the sense the pioneering experiments by Sol Spiegelman and his coworkers.[1,5] Here, we shall be concerned exclusively with this second research area: *in vitro* evolution experiments as a tool for analysing evolutionary phenomena. Work initiated with RNA molecules under conditions suitable for replication gave indeed rise to a whole new field aiming at studies of the principles of biological evolution in the laboratory. Experiments with RNA molecules fall essentially into two classes: (i) "batch procedures" where replicating molecules proliferate according to their reproductive success and (ii) selection techniques with "intervention" where criteria for survival are intoduced by the experimenter. Recently, a spin-off of these investigations became a new branch of biotechnology called "evolutionary biotechnology". It applies the knowledge of molecular evolution to desing and preparation of biopolymers for predefined purposes.

The most important prerequisite for test-tube evolution of RNA molecules is a suitable *in vitro* replication assay for nucleic acid molecules which takes care of multiplication of the molecular genotypes. In the days of Sol Spiegelman RNA replication with virus specific enzymes, so-called RNA replicases, was the only available assay for this pupose. Currently, many more amplification systems are available for nucleic acids. The most commonly used techniques combine both, template induced RNA and DNA polymerization. They are either based on reverse transcription, polymerase chain reaction (PCR),[6] and transcription or the self-sustained sequence replication (3SR) reaction.[7] In the early experiments an open system was created by means of serial-transfer[5] replenishing the replication medium through transfer of small quantities into fresh stock solution consisting of activated monomers (**ATP**, **UTP**, **GTP**, and **CTP**) and Qβ-replicase in an appropriate buffer. Efficient devices supplying the materials consumed in the replication process were conceived and built.[8] Variation is introduced into populations of molecules through mutations being several classes of replication errors: point mutations, insertions, and deletions. Rarely, also recombination events occur. Replication assays, in particular PCR, can be tuned to high error rates[9,10] thus providing sufficient diversity for selection experiments. If still more sequence variation is required, stretches of random RNA can be inserted (see, for example[11]).

Serial-transfer turned out to be an efficient tool for the design of optimal replicators under different environmental conditions. The first experiments produced RNA molecules whose replication constants were optimal under the conditions of the stock solution. In later experiments, the replication media were changed systematically for example by addition of dyes interfering with base pairing like ethidium bromide, and RNA molecules evolved which were adapted to the new conditions.[12] In recent experiments the evolutionary technique was applied to the design of RNA molecules which are resistent to cleavage by specific RNases.[8] In order to achieve that goal an automatized machinery was developed for serial transfer which allows to change experimental conditions in a precisely controlled way.

An alternative technique providing fresh replication medium continuously makes use of capillaries through which a zone of replication travels in the manner of a wave front.[13] The velocity of the front is brought into an appropriate range for observation by using a gel as medium. This setup is particularly interesting for studying the course of molecular evolution since the time coordinate is mapped into space and the history of an experiment is laid down in the inactive material behind the front of the replicating wave. It might be retrieved by analysing the gel in the capillaries.

Impressive success of molecular evolution was achived by a combination of variation and selection with intervention. The best known technique is called SELEX and is commonly used to design molecules which bind optimally to given targets.[14] The target molecules are bound convalently to a chromatographic column and suitable binders are isolated from the solution containing a great variety of candidates by retention on the column. Changing the solvent allows to produce molecules with increasing binding constants through variation and selection.[15] Other techniques based on the use of chemical tags for the identification of suitable RNA molecules were successful in changing the catalytic properties of natural ribozymes[16] as well as in the design of ribozymes with new catalytic functions.[11,17] Although a great variety of experimental results is now available, it is still very difficult if not impossible to predict optimal conditions for the design of biomolecules. Further development in the theory of molecular evolution is required for efficient planning and technological exploitation of evolution experiments. Just as chemical engineering would be doomed to fail without a solid background in chemical kinetics and material science, evolutionary biotechnology needs a comprehensive theory of (molecular) evolution for future success.

3 Theory of molecular evolution

Starting with the seminal paper of Manfred Eigen[18] a kinetic theory of molecular evolution has been conceived and developed[19,20,21] which extends conventional population genetics by considering replication and mutation explicitly as parallel chemical processes. Replication and mutation are many step polycondensation reactions that can be represented to very good approximation by a simple overall kinetics under allmost all experimental conditions.[22] In absence of RNA catalysis and without selection constraints replication-mutation kinetics leads to exponential growth of RNA genotypes. Selection constraints introduce competition into populations in the sense of Charles Darwin's natural selection. What causes the problem in modeling evolution is not the complexity of reactions but rather the hyperastronomically large number of possible genotypes which grows exponentially with chain length n (4^n for polynucleotides). Such large numbers of species are prohibitive for conventional reaction kinetics unless (simple) rules are available that allow to compute the rate and equilibrium constants of individual species from known properties of phenotypes or, in particular, from structures being the phenotypes of the RNA molecules. Statistical approaches commonly fail because of the highly complex relations between sequences and properties of phenotypes. Moreover, properties and functions of biopolymers are highly sequence specific and cannot be adequately represented by statistics. Needless to say, predefined "look-up-tables" for billions of the rate constants are not manageable. On the other hand, models based on sequence-structure relations and simple rules to derive the rate constants which are needed to describe RNA evolution are available (see forthcoming sections).

Sequences can be ordered properly by the usage of sequence space. This notion of a space of genotypes is orginally due to Sewall Wright.[23] A point is assigned to every genotype or (DNA or RNA) sequence and a distance between sequences is defined which counts the minimal number of mutations which are required to interconvert two genotypes. Restriction to point mutations simplifies the structure of genotype space, since all interconvertible sequences have the same chain length (n). The sequence space of all binary sequences

($\kappa=2$; [**G**,**C**] or [**A**,**U**]) of chain length n is a hypercube of dimension n and that of natural sequences ($\kappa=4$; [**A**,**U**,**G**,**C**]) is a straightforward generalization of this hypercube. We remark that a similar mathematically consistent notation of a metric space has been derived also for recombination[24]. Insertions and deletions complicate the conceptual frame of gentotype space but they can be included heuristically or through computer simulations. In this proposal the most of the specific examples will be restricted to a point mutation scenario for which the (generalized) hypercube applies as sequence space.

The kinetic theory of molecular evolution is primarily dealing with interplay and balance between mutation creating variability and selection reducing diversity in populations. In the limit of large populations this replication-mutation-selection scenario is described by the kinetic equations of evolution for r different molecular species,

$$\frac{dx_j}{dt} \;=\; \sum_{i=1}^{r} Q_{ij}\, F_i(\mathbf{x})\, x_i \;-\; x_j\, \Phi(\mathbf{x}) \;; \quad i,j = 1,2,\ldots,r \;, \tag{1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_r)$ and $\Phi(\mathbf{x}) = \sum_{i=1}^{r} F_i(\mathbf{x})\, x_i$ is a selection constraint that leads to constant populations size and suggests the use of normalized variables $\sum_{i=1}^{r} x_i = 1$. In mass action kinetics the functions $F_i(\mathbf{x})$ specifying reproduction of genotypes can be expanded in a power series.[b]

$$F_i(\mathbf{x}) \;=\; k_i \;+\; \sum_{\ell=1}^{r} k_{i\ell} x_\ell \;+\; \ldots \;. \tag{2}$$

The first term is by far the most important in molecular evolution since it describes template induced uncatalyzed replication. Higher order terms refer to catalyzed replication. Particularly interesting is here the second term which is linear in $F_i(\mathbf{x})$ and which gives rise to several important special cases in the limit of error-free replication ($\mathbf{Q} = \{Q_{ij}; i,j = 1,\ldots,r\} \;=\; \mathbb{I}$, the unit matrix). These ODE's describing error free replication were called replicator equations.[25] Examples are Fisher's selection equation, the hypercycle equation, and the Schlögl model.[26] For a detailed mathematical treatment of replicator equations see.[27,28,29,30,31] The replication-mutation case has been analysed in.[32,33,34] Here we shall be concerned only with uncatalyzed replication and mutation.

The simple replication-mutation-selection equation (eq.1 with $F_i(\mathbf{x}) = k_i$) is the basis of the molecular quasispecies concept[19] and has been studied in great detail.[35] A quasispecies is defined as the stationary distribution of mutants in an infinite population (see figure 1). It represents the genetic reservoir

[b] For the sake of simplicity we assume equal degradation rates or lifetimes for all genotypes This condition can be relaxed without changing the results discussed here.[18,19]

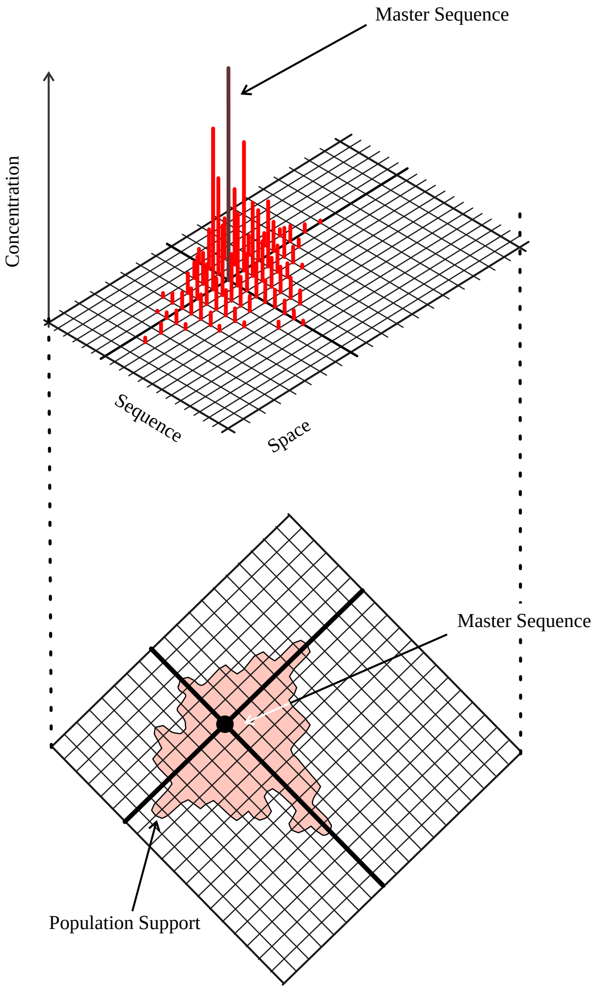

Figure 1: Molecular quasispecies in sequence space. The quasispecies is a stationary mutant distribution surrounding a (fittest and most frequent) master seqeunce. The frequencies of individual mutants in the quasispecies are determined by their fitness values and by their Hamming distances from the master. A quasispecies occupies some region in sequence space called the **population support**. In the non-stationary case the (population) support migrates through sequence scape.

of asexually replicating species like molecules in the test-tube, viruses, and bacteria. Stationary mutant distributions can be computed from an appropriately transformed linear version of the differential equation (1) by solving the corresponding eigenvalue problem.[36,37] The frequencies of individual mutants are obtained as the components of the lowest eigenvector. A typical quasispecies consists of a most frequent master sequence or master genotype $\mathbf{I}_m$ and its closely related mutants of sufficiently high fitness. Considering the quasispecies as a function of replication accuracy revealed the existence of a sharply defined error-threshold. At the critical error rate (the maximal error rate, $p_{\max}$, which is compatible with a quasispecies) the nature of the lowest eigenvalue changes abruptly from an ordered distribution around the master sequence to the uniform distribution (with all genotypes being present at equal frequencies).[c] Because of the hyperastronomically large number of genotypes a uniform distribution of mutants is incompatible with any real and hence finite population. The formal result of equal frequencies of all genotypes in the infinite population can be interpreted as an indication for random drift of real populations through sequence space in the sense of neutral evolution.[3] The critical error rate is approximated very well by the condition (where the index "m" refers to the master genotype)

$$Q_{mm} \;=\; Q_{\min} \;=\; {\sigma_m}^{-1} \quad \text{with} \quad \sigma_\mathrm{m} \;=\; \frac{\mathrm{k_m}}{\overline{\mathrm{k_{-m}}}} \quad ; \; \overline{\mathrm{k_{-m}}} \;=\; \frac{\sum_{\mathrm{i=1,i\neq m}}^{\mathrm{r}} \mathrm{k_i}}{1-\mathrm{x_m}} \; . \tag{3}$$

The existence of stationary mutant distribution in finite populations of size N requires higher accuracy of replication than in the limit of infinite population size:[38]

$$Q_{\min}(N) \;=\; Q_{\min}(\infty)\left(1+\frac{2(\sigma_m-1)}{\sqrt{N}}+\frac{2(\sigma_m-1)^2}{N}+\frac{(\sigma_m-1)^3}{(\sqrt{N})^3}+\ldots\right) \; .$$

The series expansion converges very fast alraedy for population sizes $N > 100$.

The mutation matrix $\mathbf{Q}$ is often constructed under the simplifying assumption that mutation rates are independent of the particular nucleotide exchange and the position on sequence (uniform error-rate model). Then we find for the probability that genotype $\mathbf{I}_j$ is formed as an error-copy of genotype $\mathbf{I}_i$:

$$Q_{ij} \;=\; q^{n-d_{ij}}\,(1-q)^{d_{ij}} \;=\; (1-p)^n \left(\frac{p}{1-p}\right)^{d_{ij}} \; . \tag{4}$$

[c] The existence of an error-treshold depends to some extent also on the distribution of fitness values in sequence space. There are certain classes of flat landscapes which do not support sharp thresholds and thus are characterized by smooth transitions from the quasispecies to the uniform distribution.

Herein q is the single digit accuracy and $p = 1 - q$ the error rate per site and replication. The Hamming distance between the genotypes $\mathbf{I}_i$ and $\mathbf{I}_j$ is denoted by d_{ij}. Within this model it is straightforward to compute the critical threshold value of the error rate:

$$p_{\max} \;=\; 1 \,-\, q_{\min} \;=\; 1 \,-\, \sigma_m^{-\frac{1}{n}} \;.$$

Eq.(4) allows to compute mutation probabilities for all pairs of sequences from a single parameter q and thus solves, in part, the problem to handle very large numbers of different genotypes by means of a simple rule. Still, the problem of hyperastronomically large numbers of rate constants (k_i) remains. A novel approach is thus required which allows to derive analytical expressions or algorithms for the computation of constants from known sequences and the structures derived from them.

Despite these problems in the development of a comprehensive model for biological evolution the quasispecies concept has been applied in a heuristic version to virology.[39] In particular, RNA viruses are generally characterized by low fidelity of their replicases leading to mean error numbers of 0.1 to 10 per replication. Populations of RNA viruses share high genetic diversity with those of RNA molecules replicating in test-tubes. Although virus populations live in rapidly varying environments and presumably never reach stationarity, the quasispecies concept has been adapted successfully and provides completely new insights into virus evolution which suggest to develop new antiviral strategies.

4 Modelling evolutionary dynamics

Within the last few years we conceived and developed a new concept for analysing and modeling molecular evolution (For a recent review see [40]). The overwhelming complexity is reduced through partitioning into three simpler phenomena that can be studied separately (figure 2). Population genetics of *in vitro* evolution is, in essence, described by the differential equation (1) or by suitable stochastic processes adapting it to final population sizes. For example, multitype branching processes[41] or birth-and-death processes[38] were applied successfully. Population dynamics deals with formation of new genotypes through mutation and elimination of less fit ones through selection. Details of population structure do not matter when we are interested in the migration through sequence space. It is sufficient therefore to consider only the support of the population.[d] Support dynamics describes, for example, how adaptive

[d] The **support** of a population in sequence space is the area that is covered by the actually present genotypes irrespective of their frequency (See figure 1).

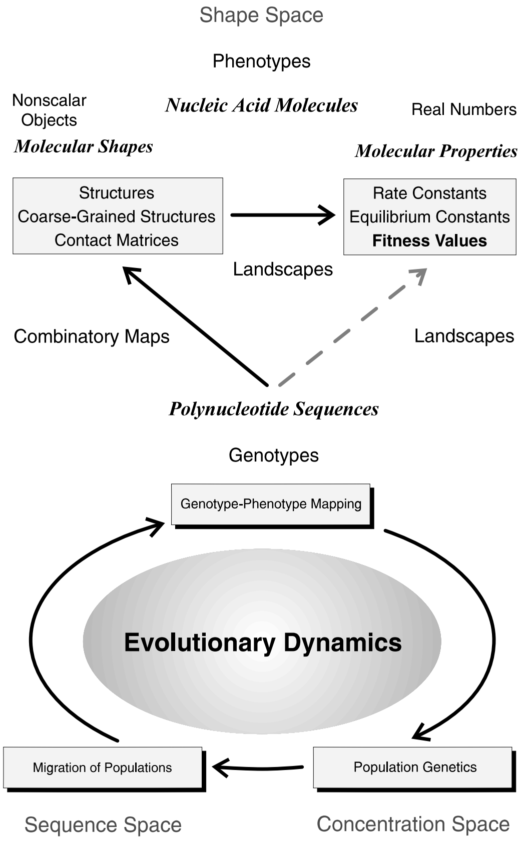

Figure 2: Evolutionary dynamics. Evolution is partitioned into three processes that can be studied separately: (i) population genetics, (ii) migration of populations, and (iii) genotype-phenotype mapping. In molecular evolution population genetics is tantamount to chemical reaction kinetics of replication, mutation and selection. Population support dynamics describes the migration of populations in in sequence space. Genotype-phenotype mapping unfolds the biological information stored in polynucleotide sequences. Two classes of mappings are distinguished: (i) combinatory maps from one genotype space into another vector space or another space of non-scalar objects and (ii) landscapes that map genotype space into the real numbers. In molecular evolution landscapes provide rate constants, equilibrium constants and other composite scalar properties of phenotypes, for example fitness values. These landscapes are commonly constructed in two steps: (i) a mapping of polynucleotide sequences into molecular structures and (ii) an evaluation of structures to yield the (scalar) molecular properties.

dynamics and random drift assist each other in evolutionary optimization. It defines the regions in sequence space from where new genotypes originate. The third phenomenon is the unfolding of phenotypes. It is the basis of the relation between genotypes and phenotypes which is understood as a mapping from sequence space into shape space. The shape space is a metric space of the phenotypes formed by all genotypes in sequence space. Distances between phenotypes or shapes can be measured in different ways (see[42] and next section). As indicated in figure 2 the three processes are linked by a cyclic relationship in the sense that genotype-phenotype mapping provides the input for population dynamics by laying down the kinetic parameters through the evaluation of phenotypes. Population genetics creates the input for support dynamics by deciding on the fate of genotypes through mutation and selection, and eventually, support dynamics closes the cycle by describing how populations migrate in sequence space and defining thereby the regions from where new genotypes come which enter genotype-phenotype mapping. Such cyclic causalities are typical for self-organisation phenomena.

Population support dynamics is dealing with the migration of populations through sequence space. The two extremes of support dynamics are: (i) adaptive walk and (ii) random drift of populations. An adaptive walk is characterized by a succession of genotypes with the restriction that each new genotype that is created and accepted in the series has to produce a phenotype with higher or at least the same fitness as the current one. On the level of populations the "no-downhill-step" condition for adaptive walks is somewhat relaxed as populations with sufficiently large population sizes can bridge narrow valleys with width of a few point mutations (see figure 7). Random drift occurs in absence of fitness differences and represents the essence of Motoo Kimura's neutral theory of evolution.[3] It can be interpreted as a diffusion process in sequence space. The only currently available analytical approach to population support dynamics is restricted to evolution on flat fitness landscapes.[43] Computer simulation of random drift has shown that growing populations may split into subpopulations.[44,45] Evolution of populations on realistic landscapes has so far only been studied by computer simulation.[45,46,47] These investigations revealed that evolutionary optimization is a combination of fast adaptive periods and slow random drift phases and thus occurs in stepwise manner with two different time scales.

In nature and in laboratory experiments, genotype-phenotype mapping is the true source of complexity.[48] Viral and bacterial phenotypes are already too complex to be studied systematically at the current state of our knowledge. The fast growing number of completely sequenced genomes, however, may change the manageability of procaryotic phenotypes. In the simplest conceivable example of a genotype-phenotype relation, *in vitro* evolution of RNA, genotype and phenotype are two features of the same molecule, sequence and structure, respectively.[1] Formation of the phenotype then is tantamount to folding the randomly coiled sequence into the stable conformation of the molecule. The structure or, in general, the phenotype links genotype and fitness since the properties which are relevant for selection are carried by the phenotype (see next section). Assignment of fitness values to genotypes is commonly done in two separate steps (A few simplified models, for example the Nk-model proposed by Stuart Kauffman[49,50] and other models related to the theory of spin glasses,[51] omit the consideration of a phenotype and assign fitness values directly to genotypes):

$$\textbf{genotype} \quad \Longrightarrow \quad \textbf{phenotype} \quad \Longrightarrow \quad \textbf{fitness} \ .$$

The first step, genotype-phenotype mapping Σ, maps one vector space onto another non-scalar space

$$\Sigma : \quad (\mathcal{S}; d_H) \quad \Longrightarrow \quad (\mathcal{Y}; \eta) \ ,$$

and has been characterized as a **combinatory map**[52,42] in order to indicate that it is no landscape in the strict sense. The set of all sequences is denoted by $\mathcal{S}$ and that of all shapes by $\mathcal{Y}$; d_H is the Hamming distance and η a distance between shapes. Fitness values are functions of the evolutionarily relevant values properties of phenotypes and, accordingly, a **fitness landscape** is a mapping from shape space into the real numbers (figure 2):

$$\Lambda : \quad (\mathcal{Y}; \eta) \quad \Longrightarrow \quad \mathbb{R}_1 \ .$$

The term "landscape" will be used here for mappings from a non-scalar space (sequence or shape space) into the real numbers, in this very general sense and irrespectively of its meaning for evolutionary dynamics.[53]

5 Genotype-phenotype mapping of RNA

Although biopolymer structures represent the simplest conceivable class of phenotypes, they are anything but easy to predict from known sequences. The precise rules which determine how three-dimensional structures are formed from sequences are not known yet. In case of RNA the empirical material consists, in essence, of roughly twenty different structures determined by x-ray crystallography and NMR-spectroscopy and thus is much poorer than the structural information available in case of proteins. RNA, however, has a meaningful level of coarse-grained structure with less detail, the so-called secondary structure which is tantamount to a list of Watson-Crick (**AU** and **GC**) and **GU** base pairs.[e] The rules of RNA secondary structure formation are sufficiently simple to allow for an analysis by means of combinatorics and other rigorous mathematical tools.[54] RNA secondary structures, on the other hand, are a fairly realistic representation of many essential features of RNA since they were used successfully for more than thirty years in biochemistry to interpret RNA reactivities and functions. In the last decade we performed a systematic study on the mapping of RNA sequences into secondary structures. These investigations are presumably dealing with the only case of a mapping from genotypes into phenotypes that is based on a realistic biophysical system and can be studied at the present state of the art. Several of its features are considered to be typical for other more complex cases in biology.

Methods to study sequence-secondary structure maps of RNA molecules are summarized in table 1. The first explorations of RNA shape space were performed by means of computer simulations of evolutionary dynamics[46,47] (see also the next section). Later on autocorrelation functions were determined for free energy landscapes and sequence-structure maps.[42,52,55] These investigations showed, for example, that landscapes derived from **GC**-only sequences are substantially more rugged than those derived from natural sequences with uniform base composition (25% **A**, 25% **U**, 25% **G**, 25% **C**). A rigorous mathematical classification of landscapes was derived by comparing different difficult optimization problems.[56,57,58]

Recent studies on the relations between RNA sequences and secondary structures used a mathematical model based on random graph theory,[59] exhaustive folding of all sequences of given chain length[60,61] as well as statistics of appropriately chosen samples.[53,55] These investigations revealed four regularities:

(i) The number of sequences exceeds the number of structures by several orders of magnitude and hence, sequence-structure maps are many to one.

(ii) Relatively few common structures are contrasted by many rare structures which usually play no role in evolution. In the limit of long chains we have almost all sequences folding in a tiny subfraction of all structures.

(iii) In order to find for any common structure at least one sequence (that folds into it under the defined criterion) one need not explore whole se-

[e] The precise definition for an acceptable secondary structure is: (i) base pairs are not allowed between neighbors in the sequences $(i, i+1)$ and (ii) if (i, j) and (k, ℓ) are two base pairs then (apart from permutations) only two arrangements along the sequence are acceptable: $(i < j < k < \ell)$ and $(i < k < \ell < j)$, respectively.

Table 1: Techniques to study mappings from RNA sequence space into the shape space of secondary structures.

	Method	Advantages	Disadvantages	Ref.
Mathematical model	Random graph theory	Analytical expressions	Limited validity of model assumptions	[59]
Exhaustive folding and enumeration	Folding algorithm and handling of large samples	Exact results	Limited to small chain lengths	[60,61]
Statistical evaluation of samples	Inverse folding and random walks in sequence space	Applicability to long sequences	Limited accuracy due to statistics	[42,52,62,63,64]
Computer simulation	Gillespie algorithm*	Focus on evolution	Scanning of small sectors in sequence space	[45,46,47,65]

* The Gillespie algorithm[66,67] is used to simulate replication and mutation in a flow reactor.

quence space. It is sufficient to search a relatively small spherical neighborhood of an arbitrarily chosen reference sequence (**shape space covering**, see figure 3 and[53]).

(iv) Common structures are characterized by a high degree of neutrality expressend by the fraction of nearest neighbors ($\bar{\lambda}$) which behave identically with respect to selection. The sets of sequences folding into them, called their preimages, form extended **neutral networks** in sequence space (figure 4).

The results derived from mappings of RNA sequences into secondary structures are of more general validity. The partitioning of structures into few common and many rare ones has been observed also with lattice models of proteins[68] and extended neutral networks of proteins were found through inverse folding using knowledge based empirical potentials of mean force.[69]

Random graph theory has been applied to model the features of the distribution of sequences in sequence space belonging to a typical neutral network.[59] This approach makes only use of the base pairing rules in secondary structures and distinguishes unpaired bases and base pairs. The generic properties of neutral networks are determined by a single parameter $\bar{\lambda}$ representing the fraction

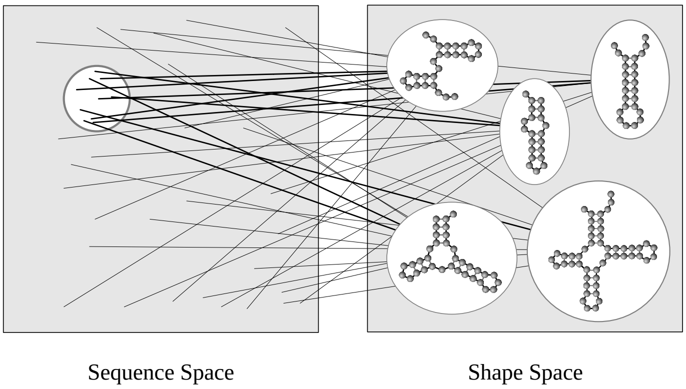

Figure 3: Shape space covering. In order to find at least one RNA sequence folding into any common structure it is not necessary to explore whole sequence space. Searching a (relatively small) spherical environment around any arbitrarily chosen reference sequence is sufficient. The radius of the covering sphere, r_{cov}, can be readily computed from properly chosen samples of structures.

of neutral neighbors in sequence space averaged over the whole network.[f] Rigorous mathematical analysis allows to derive analytical expressions for a number of relevant properties. Neutral networks are, for example, (almost always) connected and span whole sequence space when $\bar{\lambda}$ exceeds a threshold value, $\bar{\lambda} > \overline{\lambda_{cr}}(\kappa)$. Below threshold ($\bar{\lambda} < \overline{\lambda_{cr}}(\kappa)$) networks are split into components. Random graph theory predicts that there is one component, the so-called giant component which is substantially larger than the other components (figure 4). The threshold value is readily computed from the alphabet size κ:

$$\overline{\lambda_{cr}}(\kappa) \;=\; \sqrt[\kappa-1]{\frac{1}{\kappa}}\;. \tag{5}$$

The predictions of random graph theory are fulfilled well by actual neutral networks.[60,61] Exceptions can be interpreted straightforwardly in terns of structural regularities. Common structures, in general, form connected networks.

Properties (i) to (iv) of genotype-phenotype mappings are highly relevant for evolution. Restriction of searches to the highly redundant common structures and shape space covering make evolutionary optimization much simpler

[f] In the current form the model is based on a factorization of sequence space into a space of unpaired bases and a space of base pairs. Accordingly, two different $\bar{\lambda}$-values, $\bar{\lambda}_u$ and $\bar{\lambda}_p$, are used for unpaired bases and base pairs. In natural sequences the two parameters refer to two different alphabets: [**A**,**U**,**G**,**C**] and [**AU**,**UA**,**GC**,**CG**,**GU**,**UG**] with $\kappa = 4$ and $\kappa = 6$, respectively.

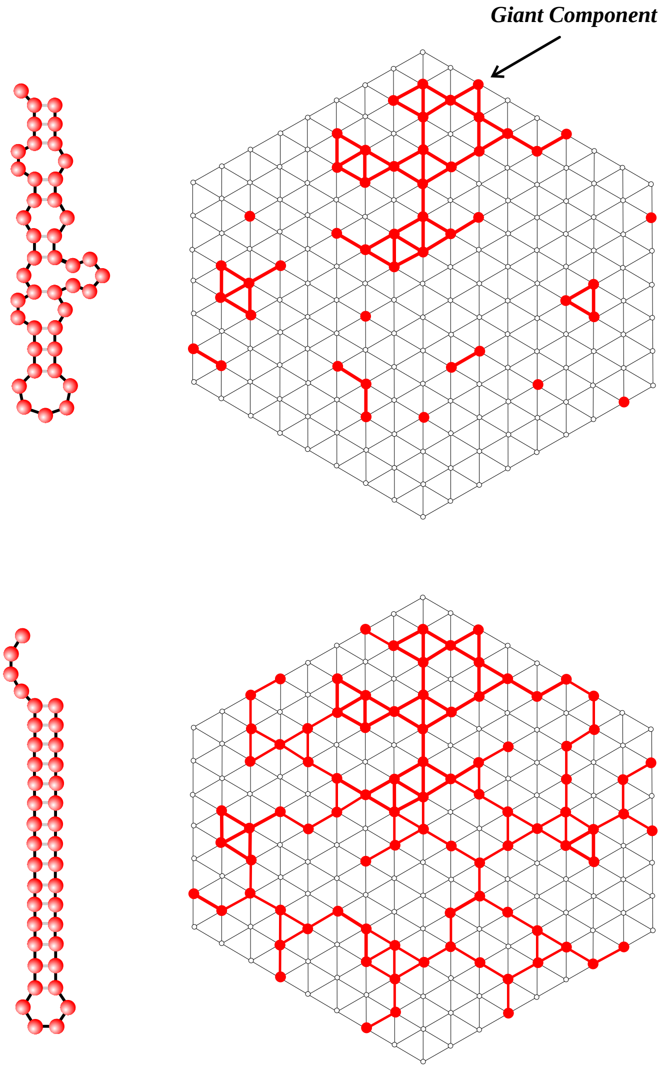

Figure 4: Neutral networks in sequence space. The lower structure forms a connected neutral network spanning whole sequence space as shown in the two-dimensional sketch. This class of network is typical for common structures. The upper part of the figure is an example of a disconnected network in sequence space which consists typically of a giant component and many small components. Connectivity of neutral networks depends on the mean fraction of neutral neighbors ($\bar{\lambda}$) of the structure in sequence space.

than previously thought and provide an explanation for the success of evolutionary biotechnology in searches were *a priori* probabilities to find a given sequence are less than 10^{-100}. In addition, shape space covering provides a powerful tool for the design of efficient protocols for searches in sequence space.[70] The existence of neutral networks is essential for the efficiency of evolutionary searches (figure 7 and[45,71,72]) since they enable populations to escape from evolutionary traps in the form of local fitness optima.

The existence of neutral networks can also be considered explicitly in the derivation of the error threshold (see section 3, eq. 3). The variables for individual genotypes forming the same phenotype are lumped together, $y_k = \sum_{i=1}^{n_k} x_i$, and thereby the following kinetic differential equation is obtained,

$$\frac{dy_k}{dt} \;=\; \sum_{j=1}^{s} \tilde{Q}_{jk}\, F_j(\mathbf{y})\, y_j \;-\; y_k\, \Phi(\mathbf{y}) \;; \quad j,k = 1,2,\ldots,s \;, \tag{6}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_s)$ and the same definitions apply as in equation (1) except for the effective mutation matrix $\tilde{\mathbf{Q}}$ whose elements are now functions of the single digit accuracy q, the degree of neutrality $\bar{\lambda}$, and the mean Hamming distances.[73] We are considering distributions of phenotypes rather than genotypes and search for the conditions of stationary phenotype distributions. The critical replication accuracy of the master phenotype becomes a function of the superiority aa well as the mean degree of neutrality:[73,74]

$$Q_{mm} \;=\; Q_{\min} \;=\; \frac{1 \,-\, \bar{\lambda}_m \sigma_m}{(1 \,-\, \bar{\lambda}_m)\, \sigma_m} \;. \tag{7}$$

The limits of the phenotypic error threshold in the $(Q, \bar{\lambda})$-plane are easily visualized: (i) the phenotypic error threshold converges to the genotypic value, $Q_{\min} = \sigma_m^{-1}$, in the limit $\bar{\lambda}_m \to 0$ and (ii) the minimal replication accuracy approaches zero in the limit $\bar{\lambda}_m \to \sigma_m^{-1}$. The second case implies that the accuracy plays no role in case the degree of neutrality is sufficiently large, i.e., when it exceeds the reciprocal value of the superiority.

6 Optimization of RNA structures

Molecular insights into evolution can be obtained by direct computer simulation of the full dynamics illustrated in figure 2. The simulated model system is based on replication and mutation in populations of RNA molecules subjected to a selection constraint through regulation of population size and genotype-phenotype mapping on the level of secondary structures. The population size is controlled by random elimination of individuals through degradation or dilution. Simulation of optimization dynamics serves in essence two purposes:

(i) The analysis of recorded data allows to give molecular interpretations of evolutionary processes which can be used for predictions and in the design of new experiments, and (ii) the results on sequence-structure mapping of RNA reported in the previous section can be tested with respect to their relevance in evolution.

The first simulations based on a realistic model of RNA structures were reported about ten years ago.[47] Like in later works populations of thousands and more RNA genotypes undergo replication and mutation and are subjected to the constraints of a flow reactor that keeps the population size N constant within fluctuations of $\sqrt{N}$-size. RNA sequences are folded to yield secondary structures.[g] The structures are then evaluated according to predefined rules in order to compute replication (k_i) and degradation rate constants (d_i). Fitness in this case is a simple function of these two quantities and the replication accuracy: $w_i = k_i Q_{ii} - d_i$. The early computer simulations[46,47] revealed, in essence, two features of evolutionary optimization: (i) the approach to the target occurs in steps, showing punctuation rather than continuity, and (ii) optimal fitness values are found with different structures strongly indicating the occurrence of selective neutrality in the evaluation of phenotypes.

More recently, simulations of this kind were used to show that evolution on the neutral network of a tRNA-structure corresponds to a diffusion process where the diffusion coefficient is proportional to the mutation rate.[45] In this simulation as well as in the computer experiment described in figures 5 and 6 degradation has been neglected and the replication rate constants (k_α) were assumed to depend on structure (independently of the sequence folding into it and thus fulfilling the neutrality condition). In particular, a (fitness) function of the kind $k_\alpha = (\delta + \eta(\alpha,\tau)/n)^{-1}$ was used, where δ is some constant, n the chain length of the RNA, and $\eta(\alpha,\tau)$ the distance between structure α and the target structure τ. Most of the evolutionarily important results, however, were found to be fairly independent of specific choices of constants and the detailed analytical expression used for the fitness function.

In our most recent works[65,77] optimization of RNA structures was studied through simulations of populations of about one thousand molecules in the flow reactor. The approach towards the target structure which happened to be a tRNA clover-leaf occurs again in steps. Periods of fast decrease in distance to the target are interrupted by long quasi-stationary phases of almost constant

[g] Folding is usually performed under a free energy minimization criterion. High-performance versions of Michael Zuker's folding algorithm[75] for sequential and parallel computing which were developed in our group[76] are applied. It should be stressed, however, that the generic results on sequence-structure maps and evolutionary dynamics presented here are fairly independent of particular folding criteria, for example maximum matching, minimum free energy or kinetic folding.[64]

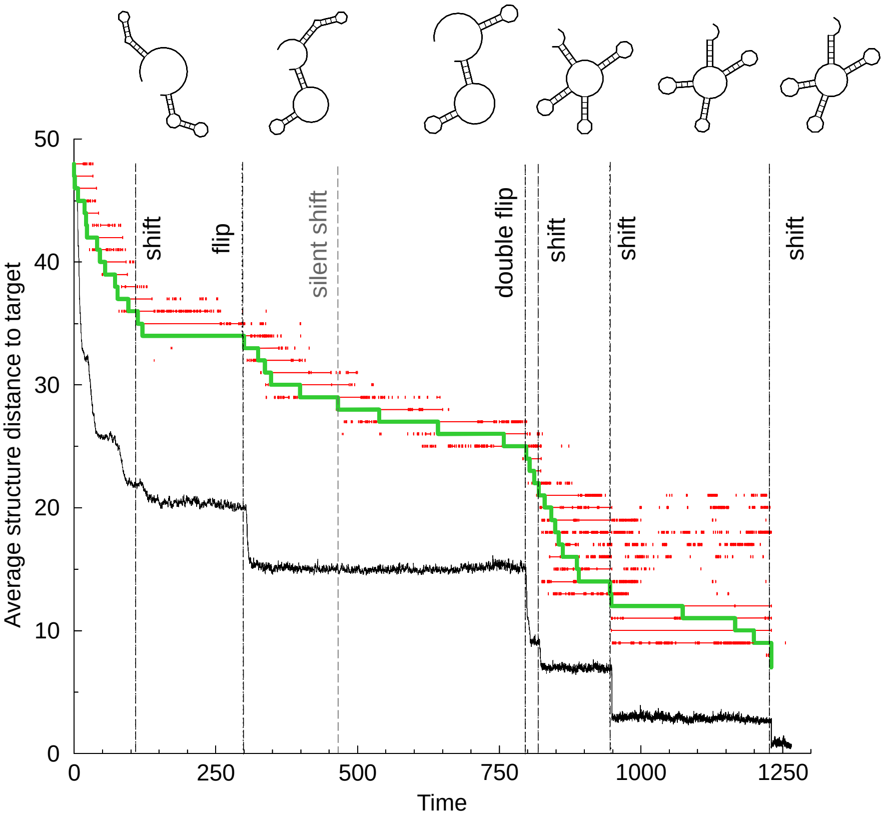

Figure 5: Transitions in a computer simulation of RNA optimization towards a tRNA shape. The figure shows how much optimization has progressed at the macro level by plotting the population average of the Hamming distance to the target structure. The fitness curve is superimposed by the relay trace showing the flow of causality from start shape to target (see text for definitions and figure 6). The approach to the target occurs in 41 steps. Seven discontinuous or major transitions are marked by vertical lines. The corresponding generalized shifts are named, and the shapes before and after the transition are shown (Except for the first standard shift to avoid congestion of the figure). All other transitions (after the first shift) are continuous in the sense that they occur within statistical neighborhoods. Horizontal intervals before and after the occurrence of a shape in the relay series indicate periods when the shape is present in the population. The flow reactor was stochastically constrained to maintain an average of 1,000 sequences of chain length $n = 76$ and the error rate was 0.001 per nucleotide.

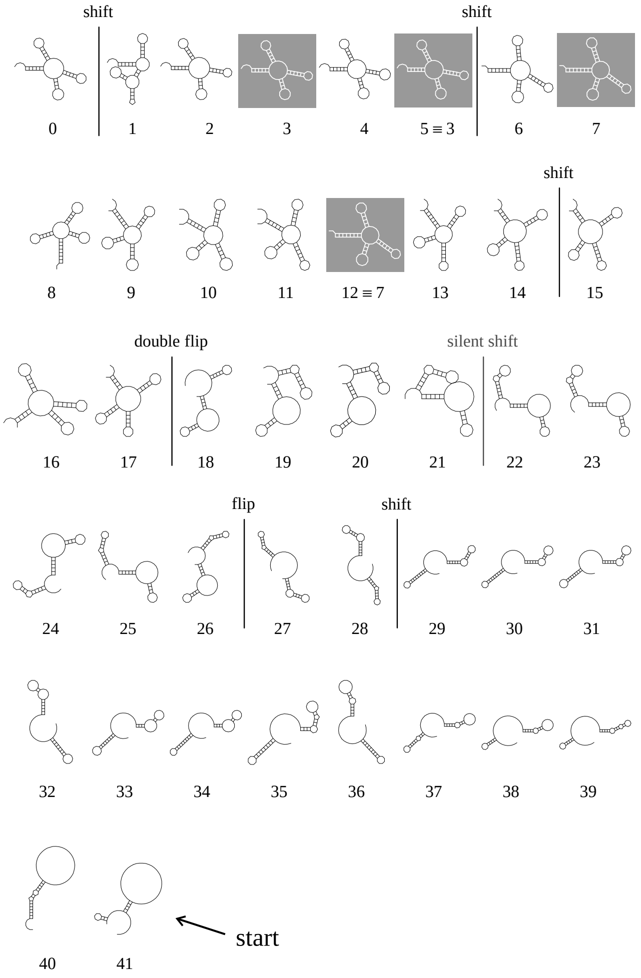

Figure 6: Relay series. The full series of 41 relay shapes derived from the computer simulation of the optimization towards a tRNA target shown in figure 5 is presented. See text for details.

average fitness (figure 5). The course of the evolutionary optimization process has been reconstructed by computing the **relay series** of phenotypes.[h] The relay series is the uninterrupted sequence of structures which eventually leads to the formation of the target structure. In computer simulations the relay series is resolved in retrospect. Starting from the end of the simulation and going back in time the population is scanned for continuous presence of the target shape and the event is determined when it appeared in the population. By this event (being a point mutation) the target shape was formed from a precursor or "parent shape". The reconstruction of the relay series is continued by determining the parent of the parent and the procedure is repeated until a shape in the initial population (at $t = 0$) has been reached. The full relay series of the computer experiment shown in figure 5 contains 41 structures (figure 6), six particularly important ones are shown on top of the figure. After an initial period of rapid improvements (which ends around time $t = 100$) the course of optimization shows a striking regularity that can be generalized to more complex systems. Transitions between structures fall into two classes:

- **continuous transitions** representing small structural changes and leading to globally frequent structures in the neighborhood of the neutral network of the intial structure and
- **discontinuous transitions** representing large stuctural changes and leading to globally rare but locally frequent structures (they are named in figure 6 according to a classification given in[65]).

Continuous transitions are minor structural changes which occur readily and involve a statistical nearness condition between neutral networks (see next section). Discontinuous transitions occur at the ends of the quasi-stationary periods (there is one exception around time $t \approx 465$ which represents a "silent" discontinuous transition that occurs in the middle of a plateau since it does not change fitness). A discontinuous transition is usually followed by a cascade of continuous transitions which are accompanied by fitness increase. Then, the population approaches the next plateau along which neutral evolution occurs at approximately constant fitness. In addition, we observe whole families of shapes appearing simultaneously at discontinuous transitions (especially instructive examples are the "shift-transitions" at $t \approx 820$ and $t \approx 950$ in figure 5). On the plateaus cycles within these families may occur in the relay series of shapes (examples are the identical shapes **3** and **5** or **7** and **12**

[h] It should be mentioned here that recordings of evolutionary histories in the sense of relay series are, in priciple, accessible through the analysis of RNA replication-mutation experiments in capillaries.[13]

in the relay series of figure 6). Two scenarios were observed in the quasi-continuous periods: population drift randomly in sequence space and genotypes vary whereas the phenotype is either constant (for example, the plateau between at $150 < t < 300$) or phenotypes change within one of the above mentioned families. The drift continues until a point in sequence space is reached where a fitness-improving discontinuous transition is locally frequent. Repeated optimization runs from identical initial populations towards the same target structure but with different "seeds" of the random number generator proceed through different intermediates. Gross features of the simulations, however, turned out to be fairly reproducible. These are, for example, the numbers of steps, the overall features of intermediate shapes, the attainability of shapes in sequence space as well as the above reported regularities in the relay series.

A remarkable difference has been observed between **AUGC** and **GC** sequences: most of the individual runs with populations of **AUGC** sequences heading for a tRNA target shape reached the goal within some onethousand-fivehundred time units. We tried also to search in the same way for **GC**-only sequences that form tRNA structures. Although such sequences were obtained through inverse folding and thus are known to exist, none of the computer simulations with a population of one thousand individuals was successful within several thousand time units. The simulations thus confirm what has already been conjectured from the shorter correlation lengths of **GC**-only landscapes: **GC** sequences form more rugged landscapes and evolutionary optimization on them is more difficult, accordingly.

The course of evolutionary optimization on realistic landscapes is sketched in figure 7. Ruggedness of fitness landscapes lacking neutrality causes adaptive walks of populations to end on nearby local optima. Neutral networks mediate between different regions in sequence space since populations migrate on them by random drift. Optimization on landscapes with sufficiently high degree of neutrality occurs on two time scales: fast periods containing cascades of adaptive changes are interrupted by long quasi-stationary phases of neutral evolution during which populations drift randomly on neutral networks until they reach a local neighborhood that sustains the next major transition.

7 Statistical topology and evolution

The RNA model and the evolutionary dynamics derived from it inspired the development of a statistical notion of nearness in genotype space that can be formulated as a kind of **statistical topology**.[65] It allows straighforward generalization to other evolutionary systems. An evolutionarily relevant notion

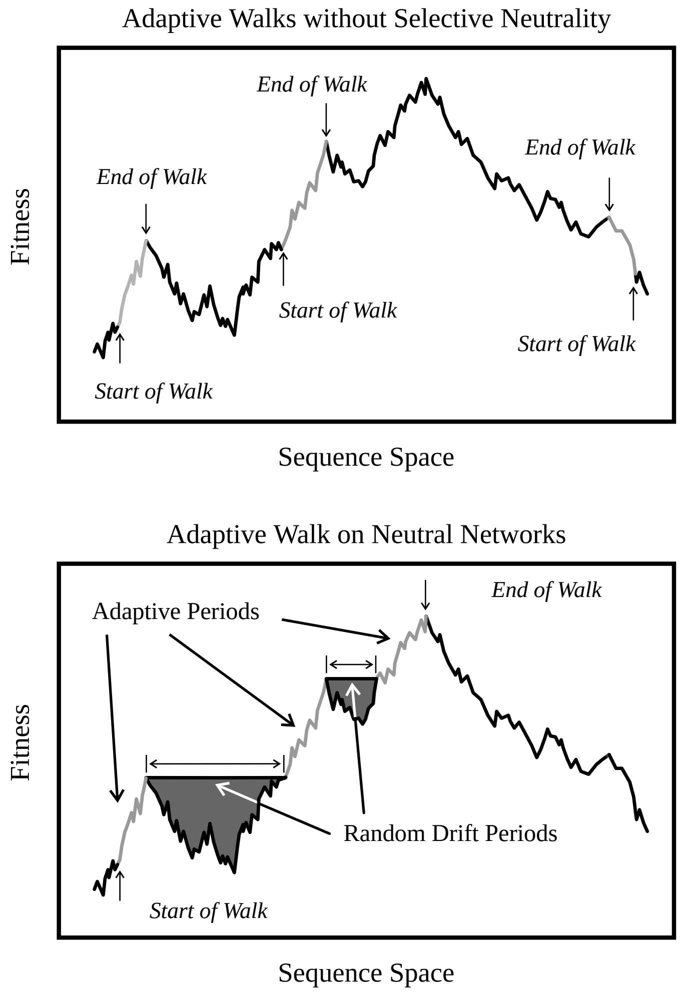

Figure 7: A sketch of optimization in sequence space through adaptive walks of populations. Adaptive walks allow to choose the next step arbitrarily from all directions where fitness is (locally) non-decreasing. Because of their quasispecies-like mutant distributions populations can bridge over narrow valleys with widths of a few point mutations. In absence of selective neutrality (upper part) they are, however, unable to span larger Hamming distances with low fitness intermediates. Hence, adaptive walks will end on one of the nearest major fitness peak. Populations on rugged landscapes with sufficiently high degree of neutrality form extended neutral networks and evolve by a combination of adaptive walks and random drift at essentially constant fitness along the network (lower part). Eventually, populations may reach the global maximum of the fitness landscape.

of **nearness** is obtained by restricting the property to be **near** to frequent occurence of structures in the neighborhood of neutral sets. Neighborhood frequency is computed by counting the shapes in all one-error (Hamming-distance-one) neighborhoods of the genotypes belonging to the network and then forming the average. Inspection of the frequency of occurence[i] allows to identify globally near phenotypes.

Shapes in the statistical nearness relation need not be commutable: shape α is near shape β does not imply that β is near α. This paradox is easily resolved by considering neutral networks of largely different sizes: the smaller network (β) may have the larger one (α) as a frequent neighbor; at the same time, however, it may occupy only a negligibly small fraction of the positions in the neighborhood of the larger network and thus β is not near α. Precisely this situation is found with tRNA's and structures derived from them by opening the terminal stack (three-hairpin-RNA): the tRNA forms the smaller network and is not near the three-hairpin-RNA whereas the three-hairpin-RNA is a frequent neighbor of the tRNA. Transitions between globally near phenotypes are continuous and occur readily. Discontinuous transitions occur between globally distant phenotypes. They are initiated by special genotypes which meet the sequence requirement that the major changes can occur through a single point mutation. Preliminary inspection of discontinuous transitions in the RNA model has shown that they are indeed locally frequent.

The nearness property of phenotypes is not restricted to RNA secondary structures. It is merely based on a sufficiently large degree of neutrality and genotype-phenotype relations which fulfil the conditions listed in section 5. Then, evolution will always appear as a sequence of continuous and discontinuous transitions where the latter depend on special genotype requirements. The role of neutral evolution is to search for these special genotypes through random drift.

Acknowledgments

The work reported here has been supported financially by the Austrian *Fonds zur Förderung der wissenschaftlichen Forschung* (Projects No.9942-PHY, No.10578-MAT, and No.11065-CHE), by the Commission of the European Community (CEC Contract Study PSS*0884), and by the Santa Fe Institute. Fruitful discussions with Drs. Walter Fontana, Christian Forst and Peter Stadler are gratefully acknowledged.

[i] Partitioning of shapes in the one-error neighborhood of neutral networks is most easily done with the help of log(rank)-log(frequency) plots commonly used to search sets for the validity of Zipf's law.[78] All neighborhoods computed so far showed a few (approximately 10 to 20) frequent shapes which are clearly set off against the other less frequent structures.[65]

References

1. S. Spiegelman. *Quart. Rev. Biophys.*, 4:213–253, 1971.
2. M. O. Dayhoff and W. C. Barker. Mechanisms in molecular evolution: Examples. In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure, Vol. 5*, pages 41–45. Natl. Biomed. Res. Found., Silver Spring, MD, 1972.
3. M. Kimura. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK, 1983.
4. J. L. King and T. H. Jukes. *Science*, 164:788–798, 1969.
5. D. R. Mills, R. L. Peterson, and S. Spiegelman. *Proc. Natl. Acad. Sci. USA*, 58:217–224, 1967.
6. K. Mullis, F. Faloona, S.Scharf, R. Saiki, G. Horn, and H. Ehrlich. *Cold Spring Harbor Symp. Quant. Biol.*, 51:263–273, 1986.
7. E. Fahy, D. Y. Kwoh, and T. R. Gingeras. *PCR Methods Appl.*, 1:25–33, 1991.
8. G. Strunk and T. Ederhof. *Biophys.Chem.*, 66:193–202, 1997.
9. R. C. Cadwell and G. F. Joyce. *PCR Methods Appl.*, 2:28–33, 1992.
10. K. A. Eckert and T. A. Kunkel. *PCR Methods Appl.*, 1:17–24, 1991.
11. D. P. Bartel and J. W. Szostak. *Science*, 261:1411–1418, 1993.
12. F. R. Kramer, D. R. Mills, P. E. Cole, T. Nishihara, and S. Spiegelman. *J. Mol. Biol.*, 89:719–736, 1974.
13. G. J. Bauer, J. S. McCaskill, and H. Otten. *Proc. Natl. Acad. Sci. USA*, 86:7937–7941, 1989.
14. C. Tuerk and L. Gold. *Science*, 249:505–510, 1990.
15. A. D. Ellington. *Current Biology*, 4:427–429, 1994.
16. A. A. Beaudry and G. F. Joyce. *Science*, 257:635–641, 1992.
17. E. H. Ekland, J. W. Szostak, and D. P. Bartel. *Science*, 269:364–370, 1995.
18. M. Eigen. *Naturwissenschaften*, 58:465–523, 1971.
19. M. Eigen and P. Schuster. *Naturwissenschaften*, 64:541–565, 1977.
20. M. Eigen and P. Schuster. *Naturwissenschaften*, 65:7–41, 1978.
21. M. Eigen and P. Schuster. *Naturwissenschaften*, 65:341–369, 1978.
22. C. K. Biebricher and M. Eigen. Kinetics of RNA replication by Qβ replicase. In E. Domingo, J. J. Holland, and P. Ahlquist, editors, *RNA Genetics. Vol.I: RNA Directed Virus Replication*, pages 1–21. CRC Press, Boca Raton, FL, 1988.
23. S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In D. F. Jones, editor, *Int. Proceedings of the Sixth International Congress on Genetics*, volume 1, pages 356–366, 1932.

24. P. F. Stadler and G. Wagner. *Evol. Comp.*, 5:241–275, 1998.
25. P. Schuster and K. Sigmund. *J. Theor. Biol.*, 100:533–538, 1983.
26. J. Hofbauer and K. Sigmund. *The Theory of Evolution and Dynamical Systems.* Cambridge University Press, Cambridge UK, 1988.
27. J. Hofbauer, P. Schuster, and K. Sigmund. *J.Math.Biol.*, 11:155–168, 1981.
28. J. Hofbauer, P. Schuster, K. Sigmund, and R. Wolff. *SIAM J.Appl.Math.*, 38:282–304, 1980.
29. P. Schuster, K. Sigmund, and R. Wolff. *Bull.Math.Biol.*, 40:743–769, 1977.
30. P. Schuster, K. Sigmund, and R. Wolff. *J.Diff.Equ*, 32:357–368, 1979.
31. P. Schuster, K. Sigmund, and R. Wolff. *J.Math.Anal.Appl.*, 78:88–112, 1980.
32. P. F. Stadler, W. Schnabl, C. Forst, and P. Schuster. *Bull.Math.Biol.*, 57:21–61, 1995.
33. P. F. Stadler and P. Schuster. *Bull.Math.Biol.*, 52:485–508, 1990.
34. P. F. Stadler and P. Schuster. *J.Math.Biol.*, 30:597–632, 1992.
35. M. Eigen, J. McCaskill, and P. Schuster. *Adv. Chem. Phys.*, 75:149 – 263, 1989.
36. B. L. Jones, R. H. Enns, and S. S. Rangnekar. *Bull. Math. Biol.*, 38:12–28, 1975.
37. C. J. Thompson and J. L. McBride. *Math. Biosc.*, 21:127–142, 1974.
38. M. Nowak and P. Schuster. *J. Theor. Biol.*, 137:375–395, 1989.
39. E. Domingo and J. J. Holland. *Ann.Rev.Microbiol.*, 51:151–178, 1997.
40. P. Schuster. *Physica D*, 107:351–365, 1997.
41. L. Demetrius, P. Schuster, and K. Sigmund. *Bull. Math. Biol.*, 47:239–262, 1985.
42. W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster. *Biopolymers*, 33:1389–1404, 1993.
43. B. Derrida and L. Peliti. *Bull. Math. Biol.*, 53:355–382, 1991.
44. P. G. Higgs and B. Derrida. *J. Mol. Evol.*, 35:454–465, 1992.
45. M. A. Huynen, P. F.Stadler, and W. Fontana. *Proc. Natl. Acad. Sci. USA*, 93:397–401, 1996.
46. W. Fontana, W. Schnabl, and P. Schuster. *Phys. Rev. A*, 40:3301–3321, 1989.
47. W. Fontana and P. Schuster. *Biophys. Chem.*, 26:123–147, 1987.
48. P. Schuster. *Complexity*, 2:22–30, 1996.
49. S. A. Kauffman and S. Levine. *J. Theor. Biol.*, 128:11–45, 1987.
50. S. A. Kauffman and E. D. Weinberger. *J. Theor. Biol.*, 141:211–245, 1989.

51. C. Amitrano, L. Peliti, and M. Saber. A spin-glass model of evolution. In A. S. Perelson and S. A. Kauffman, editors, *Molecular Evolution on Rugged Landscapes*, volume IX of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 27–38. Addison-Wesley Publ. Co., Redwood City, CA, 1991.
52. W. Fontana, T. Griesmacher, W. Schnabl, P. F. Stadler, and P. Schuster. *Mh. Chem.*, 122:795–819, 1991.
53. P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. *Proc.Roy.Soc.(London)B*, 255:279–284, 1994.
54. M. S. Waterman. *Introduction to Computational Biology. Maps, Sequences, and Genomes.* Chapman & Hall, London, 1995.
55. W. Fontana, P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, and P. Schuster. *Phys. Rev. E*, 47:2083–2099, 1993.
56. P. Schuster and P. F. Stadler. *Computers Chem.*, 18:295–314, 1994.
57. P. F. Stadler. Towards a theory of landscapes. In R. Lopéz-Peña, R. Capovilla, R. García-Pelayo, H. Waelbroeck, and F. Zertuche, editors, *Complex Systems and Binary Networks*, pages 77–163, Berlin, New York, 1995. Springer Verlag.
58. P. F. Stadler. *J. Math. Chem.*, 20:1–45, 1996.
59. C. Reidys, P. F. Stadler, and P. Schuster. *Bull. Math. Biol.*, 59:339–397, 1997.
60. W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. *Mh.Chem.*, 127:355–374, 1996.
61. W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. *Mh.Chem.*, 127:375–389, 1996.
62. S. Bonhoeffer and P. F. Stadler. *J. Theor. Biol.*, 164:359–372, 1993.
63. M. Tacker, W. Fontana, P. F. Stadler, and P. Schuster. *Eur. Biophys. J.*, 23:29–38, 1994.
64. M. Tacker, P. F. Stadler, E. G. Bornberg-Bauer, I. L. Hofacker, and P. Schuster. *Eur.Biophys.J.*, 25:115–130, 1996.
65. W. Fontana and P. Schuster. *J.Theor.Biol.*, 1998.
66. D. T. Gillespie. *J. Comp. Phys.*, 22:403–434, 1976.
67. D. T. Gillespie. *J. Phys. Chem.*, 81:2340–2361, 1977.
68. H. Li, R. Helling, C. Tang, and N. Wingreen. *Science*, 273:666–669, 1996.
69. A. Babajide, I. L. Hofacker, M. J. Sippl, and P. F. Stadler. *Folding & Design*, 2:261–269, 1997.
70. P. Schuster. *Journal of Biotechnology*, 41:239–257, 1995.
71. M. A. Huynen. *J. Mol. Evol.*, 43:165–169, 1996.

72. P. Schuster. The role of neutral mutations in the evolution of RNA molecules. In S. Suhai, editor, *Theoretical and Computational Methods in Genome Research*, pages 287–302, New York, 1997. Plenum Press.

73. C. Reidys, C. V. Forst, and P. Schuster. Replication and mutation on neutral networks. Submitted, 1998.

74. P. Schuster. *Biophys. Chem.*, 66:75–110, 1997.

75. M. Zuker and D. Sankoff. *Bull. Math. Biol.*, 46:591–621, 1984.

76. I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. *Mh. Chem.*, 125:167–188, 1994.

77. W. Fontana and P. Schuster. Stepping through phenotype space. On the nature of transitions. Submitted, 1998.

78. G. Zipf. *Human Behaviour and the Principle of Least Effort.* Addison-Wesley, Reading, MA, 1949.