CNLS 16th Annual International Conference: Los Alamos, NM, May 13-16, 1996
LANDSCAPE PARADIGMS IN PHYSICS AND BIOLOGY

LANDSCAPES AND MOLECULAR EVOLUTION

Peter SCHUSTER*
Santa Fe Institute, Santa Fe, New Mexico 87501, USA,
Institut für Molekulare Biotechnologie, D-07708 Jena, Germany, and
Institut für Theoretische Chemie der Universität Wien, A-1090 Wien, Austria

Evolution of RNA molecules *in vitro* is visualized as a hill-climbing process on a fitness landscape that can be derived from molecular properties and functions. The optimization process is shaped by a high degree of redundance in sequence-to-structure mappings: there are many more sequences than structures and sequences folding into the same structure are (almost) randomly distributed in sequence space. Two consequences of this redundance are important for evolution: **shape space covering** by small connected regions in sequence space and the existence of extended **neutral networks**. Both results together explain how nature can fast and efficiently find solutions to complex optimization problems by trial and error while the number of possible genotypes exceeds all imagination. In the presence of neutral networks populations avoid being caught in evolutionary traps and eventually reach the global optimum through a composite dynamics of adaptive walks **and** random drift. Results derived from mathematical analysis are confronted with the results of computer simulation and available experimental data.

1. Evolution and landscapes

Sewall Wright [72] created the metaphor of hill-climbing on a fitness landscape for Darwinian evolution. Optimization then corresponds to an adaptive walk, i.e., to a process that allows to choose the direction for the next step at random from all directions along which fitness does not decrease. Stationary states of populations correspond to local optima of the fitness landscape. Evolution is seen as a series of transitions between optima with increasing fitness values. Wright's metaphor saw a recent revival when sufficiently simple models of fitness landscapes became available [1, 41]. These models are based on spin glass theory [63, 66] or closely related to it like Kauffman's Nk model [42]. Evolution of RNA molecules has been studied by more realistic models that deal explicitly with molecular structures obtained from folding RNA sequences [23, 24]. Fitness values serving as input parameters for evolutionary dynamics were derived through evaluation of the structures. The complexity of RNA fitness landscapes originates from conflicting consequences of structural changes that are reminiscent of "frustration" in the theory of spin glasses [2]. Fitness in the Darwinian sense means numerous progeny. Large fitness values can be achieved by high replication rate or

---

*Correspondence should be sent to: Institut für Molekulare Biotechnologie, PF 100813, D-07708 Jena, Germany. E-Mail: `pks@imb-jena.de`, Phone: +49 (3641) 65 6444, Fax: +49 (3641) 65 6446

longevity, i.e., small degradation rate. Structures that are optimal for replication in the RNA model are also easily degraded and hence there is no simple optimal solution with high replication and low degradation rate.

Biological evolution is a highly sophisticated dynamical phenomenon, and its complexity is often confusing. For the purpose of analysis and better understanding it is partitioned into the three simpler processes shown in figure 1 [56]: population dynamics, (population) support dynamics and genotype-phenotype mapping. The three processes are properly visualized in three abstract metric spaces:
(1) the *concentration space* of biochemical reaction kinetics,
(2) the *sequence space* of polynucleotide sequences, and
(3) the *shape space* of biopolymer structures.
Kinetics of chemical reactions or changes in populations are recorded in concentration space, the conventional space of chemical reaction kinetics. It was formalized and put into precise mathematical terms by Martin Feinberg [19]. Variables count numbers of particles, molecules, virions, cells or organisms, for example, and display their changes over time. A population of N individuals distributed over m variants is denoted by $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ with $\sum_{i=1}^{m} x_i = N$. Population dynamics is commonly described by differential equations,

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})\,; \quad \mathbf{f}(\mathbf{x}) = \Big(f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\Big)\,, \tag{1}$$

by difference equations in case of synchronized generations or by a stochastic process in case of small particle numbers. Concentration space is restricted to the classes of genotypes actually present. When a mutant is produced, a new variable appears in concentration space; when a variant dies out, the corresponding variable disappears. The number of variables (m) matches the number of currently existing genotype classes. As a well known example we mention the selection mutation equation for asexually reproducing individuals introduced by Eigen [13]:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = x_i \Big(k_i Q_{ii} - d_i - \Phi(\mathbf{x})\Big) + \sum_{j \neq i} k_j Q_{ji} x_j \;; \quad i = 1, 2, \ldots, m \tag{2}$$

Replication and degradation rate constants are denoted by k_i and d_i, respectively, replication accuracies and mutation frequencies are contained in the (bistochastic) matrix $\mathbf{Q} \doteq \{Q_{ij}\}$ ($\sum_{j=1}^{m} Q_{ij} = 1$, $\sum_{i=1}^{m} Q_{ij} = 1$), and $\Phi(\mathbf{x}) = \sum_{i=1}^{m} (k_i - d_i) x_i = \sum_{i=1}^{m} e_i x_i$ is the mean excess production of the population. Population variables are assumed to be normalized and thus the physically meaningful range of variables is confined to the concentration simplex: $S_m \doteq \big\{x_i \geq 0 \,\forall\, i = 1, \ldots, m;\ \sum_{i=1}^{m} x_i = 1\big\}$.

The sequence space is a metric point space $(\mathcal{S}; h)$ containing all possible κ^n genotypes of given chain length n. By κ we denote the size of the polynucleotide alphabet (**AUGC**:

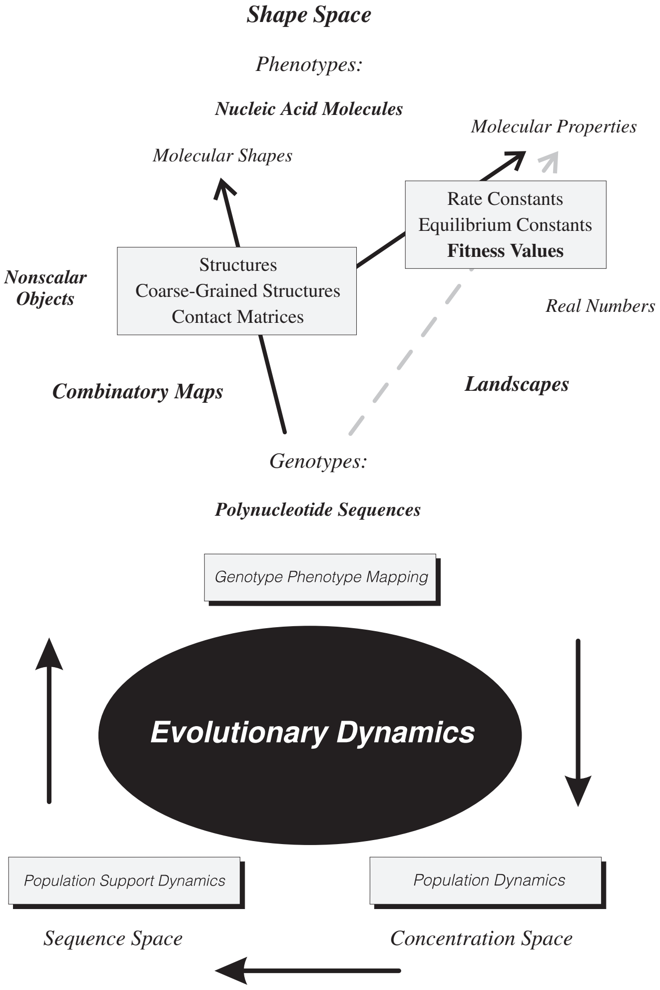


Figure 1: Evolutionary dynamics. *In vitro* evolution of biomolecules is partitioned into three simpler processes: (i) population dynamics, (ii) population support dynamics, and (iii) genotype-phenotype mapping. Population dynamics is tantamount to chemical reaction kinetics of replication, mutation and selection. Population support dynamics describes the migration of populations in the space of genotypes (see section 6). Genotype-phenotype mapping unfolds biological information stored in polynucleotide sequences. Two classes of mappings are distinguished: combinatory maps from one genotype space into another vector space or another space of non-scalar objects and landscapes that map genotype space into the real numbers. In molecular evolution landscapes provide rate constants, equilibrium constants and other more complex scalar properties of phenotypes, fitness values for example. These landscapes are commonly constructed in two steps: (i) a mapping of polynucleotide sequences into molecular structures and (ii) an evolution of structures to yield the (scalar) molecular properties.

$\kappa = 4$, and **GC**: $\kappa = 2$). Distances in sequence space are expressed by the minimum number of single nucleotide exchanges or point mutations converting one sequence into another. This distance, called Hamming distance h [27], induces a metric in sequence space. Connecting all pairs of binary sequences of chain length n with Hamming distance one by straight lines one obtains a hypercube of dimension n. Figure 2 shows how the sequence space in the four letter alphabet can be related to the corresponding binary sequence space. The notion of sequence space is useful for modeling mutation frequencies because genotypes are grouped according to error classes. The uniform error rate model [15, 16], for example, allows to express all mutation rates in terms of only one tunable parameter, the single digit accuracy q (or the error rate $p = 1 - q$), the chain length n and the Hamming distance between template and mutant h_{ij}:

$$Q_{ij} \;=\; q^n \left(\frac{1-q}{q}\right)^{h_{ij}} \;=\; q^n \left(\frac{p}{1-p}\right)^{h_{ij}} \; . \qquad (3)$$

Here the accuracy of replication ($Q_{ii} = q^n$) is independent of the nucleotide sequence.

The shape space contains all phenotypes that are formed by processing genotypes in a given context. In molecular evolution phenotypes are tantamount to spatial structures of biopolymers [65]. Meaningful notions of structure, however, depend on the context: successful descriptions of active sites of enzymes (or ribozymes) or specific binding sites for regulatory elements on DNA, for example, require high precision at atomic resolution as provided by x-ray crystallography whereas studies on phylogenetic conservation of structures can be done much better on the coarse-grained level of *ribbon* diagrams. Some levels of coarse-graining, secondary structures of RNA molecules, for example, are not only of physical relevance but also suitable for mathematical modeling. We assume that phenotypes are elements of some metric space $(\mathcal{Y};\eta)$. In other words, the existence of a distance, η, measuring the relatedness of phenotypes is postulated (although it is not explicitly required for most applications).

The assignment of fitness values to individual genotypes is done in two separate steps (several models, Kauffman's Nk model [42] for example, omit the consideration of a phenotype and assign fitness values directly to genotypes):

$$\boldsymbol{genotype} \;\Longrightarrow\; \boldsymbol{phenotype} \;\Longrightarrow\; \textbf{fitness}\,.$$

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

$$\Sigma : \quad \big(\mathcal{S};h\big) \;\Longrightarrow\; \big(\mathcal{Y};\eta\big) \; , \qquad (4)$$

and has been characterized as a combinatory map [21, 22] in order to indicate that it is no proper landscape. Fitness values are functions of the evolutionarily relevant properties of phenotypes and a fitness landscape, accordingly, is a mapping from shape space into the real numbers (fig.1):

$$\Lambda : \quad \big(\mathcal{Y};\eta\big) \;\Longrightarrow\; \mathbb{R}_1 \; . \qquad (5)$$

The term "landscape" in this general sense will be used here for a mapping from sequence space into the real numbers irrespectively of its meaning for evolutionary dynamics [60].

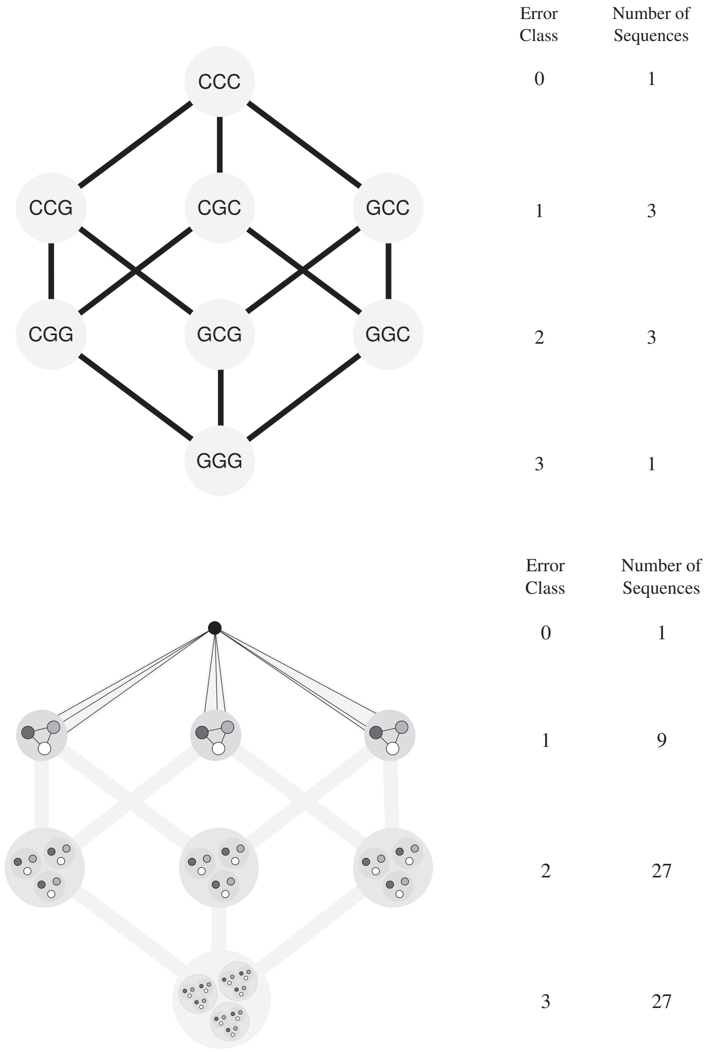


Figure 2: Sequence space. Sequences are represented by points in sequence space. Edges connect sequences of Hamming distance one, i.e., sequences that differ in a single position only. The index of the error class counts the Hamming distance from the reference sequence. The upper part shows the sequence space of **GC**-only sequences of chain length n=3. It has the form of a cube in ordinary three-dimensional space. In general, the sequence spaces of binary sequences are hypercubes of dimension n. The lower part contains the sequence space of natural (**AUGC**) sequences of chain length n=3. Note, that every base can be replaced by three other bases, for example, **A**⟶**U**, **A**⟶**G**, and **A**⟶**C**. The different replacements are coded in different grey levels. Edges are only shown for the error classes 0 and 1.

2. Gradient dynamics

In case of optimization through Darwin's principle of variation and selection leading to "survival of the fittest" population dynamics consists simply of a monotonous approach towards a steady state consisting of a homogeneous population of the fittest genotype. This phenomenon was formulated in mathematical terms through Fisher's fundamental theorem of natural selection [20] which states that the mean fitness of a population, $\bar{\varphi}(t) = \sum_{i=1}^{m} \varphi_i x_i(t)$, is steadily increasing. The theorem is tantamount to saying that mean fitness is a Ljapounov function of the dynamical system (1). In the simplest cases the r.h.s. of the differential equation describing population dynamics represents a gradient

$$\dot{\mathbf{x}} \;=\; \operatorname{grad} V(\mathbf{x}) \;, \tag{6}$$

implying that all trajectories of the ODE are (locally) perpendicular to the constant level sets of $V(\mathbf{x})$. In other cases a Riemannian metric of sequence space has been introduced (instead of the conventional Euclidean metric) in order to transform (1) into a generalized gradient system [32, 55, 62, 64]. In particular, Shahshahani [62] defined such a gradient ($Grad$) based on the scalar product $|\mathbf{x},\mathbf{y}|_{\mathbf{z}} = \sum_{i=1}^{m} x_i y_i \,/ z_i$ whose application turns the selection equation (equ.2 with $\mathbf{Q} = \mathbb{I}$, the unit matrix) into a gradient: $\dot{\mathbf{x}} = Grad\,\Phi(\mathbf{x})$. It is worth noticing that the introduction of non-zero mutation rates restricts optimization of $\Phi(\mathbf{x})$ to parts of the concentration simplex. In other parts of concentration space $\Phi\big(\mathbf{x}(t)\big)$ may be monotonously decreasing or even a non-monotonous function of time [55, 61].

The gradient form of the selection equation is indeed the mathematical basis for considering evolving populations as walkers climbing uphill on a (rugged) fitness landscape in sequence space [40, 72] (see section 6). Many natural processes, however, interfere with pure selection and restrict the hill-climbing scenario to parts of the concentration simplex or make it entirely obsolete. Examples are mutation, recombination and frequency dependent reproduction. The latter case has been studied extensively by means of replicator equations (for a survey see e.g. Hofbauer & Sigmund [33]). Population dynamics may indeed be as complex as deterministic chaos. If one focusses on replicator dynamics [59] oscillations of concentrations [17], deterministic chaos in homogeneous solutions [53], and spiral waves in spatially resolved systems [6] have been reported.

3. Evolution of RNA molecules

Experimental studies on biological evolution are facing three tantalizing problems: (i) evolutionary phenomena like adaptation need tenthousands to millions of generations in order to become observable in populations and the times required for experiments much too long, (ii) combinatorial explosion leads to numbers of possible genotypes that exceed all imaginations, and (iii) relations between genotypes and phenotypes are so complex that realistic modeling is impossible. Even in the case of fast replicating bacteria with generation times of approximately one hour tenthousand generations require more than a year and evolution experiments would last several years (Patient microbiologists, for example, have just arrived at some tenthousand generations of *Escherichia coli* under controlled conditions [46]). The numbers of possible sequences are prohibitive for systematic studies: viroid genoms are about threehundred nucleotides long and this implies already 4×10^{180} different RNA sequences of this chain length. The complexity of the relation between genotypes and phenotypes is even more discouraging: to predict the spatial structure of a single biopolymer molecule from its sequence is a very hard (and still unsolved) problem, and the simplest bacteria consist of several thousand different protein and nucleic acids molecules.

The first *in vitro* evolution experiments were performed in the sixties by Sol Spiegelman (for a review see [65]): replication rate constants of RNA from the bacteriophage $Q\beta$ were optimized in serial transfer experiments in suitable replication media. These studies demonstrated that evolutionary phenomena are not restricted to the existence of cellular life: molecules capable of reproduction and mutation fulfil the prerequisites for Darwin's principle and behave like asexually replicating individuals (as far as selection and adaptation to environmental conditions are concerned). About the same time Manfred Eigen [13] developed a theoretical frame for molecular evolution which had its roots in chemical reaction kinetics. *In vitro* evolution of RNA molecules circumvents the three problems indicated above: (i) generation times can be reduced to a few seconds under favorable conditions and evolutionary phenomena become observable in a few days, (ii) many genotypes form the same phenotype and are thus selectively neutral in the sense of Motoo Kimura [43], and (iii) genotype and phenotype are two properties of the same molecule, sequence and spatial structure, respectively. Evolution of RNA in the test-tube represents the simplest conceivable system that allows to study adaptation in the laboratory.

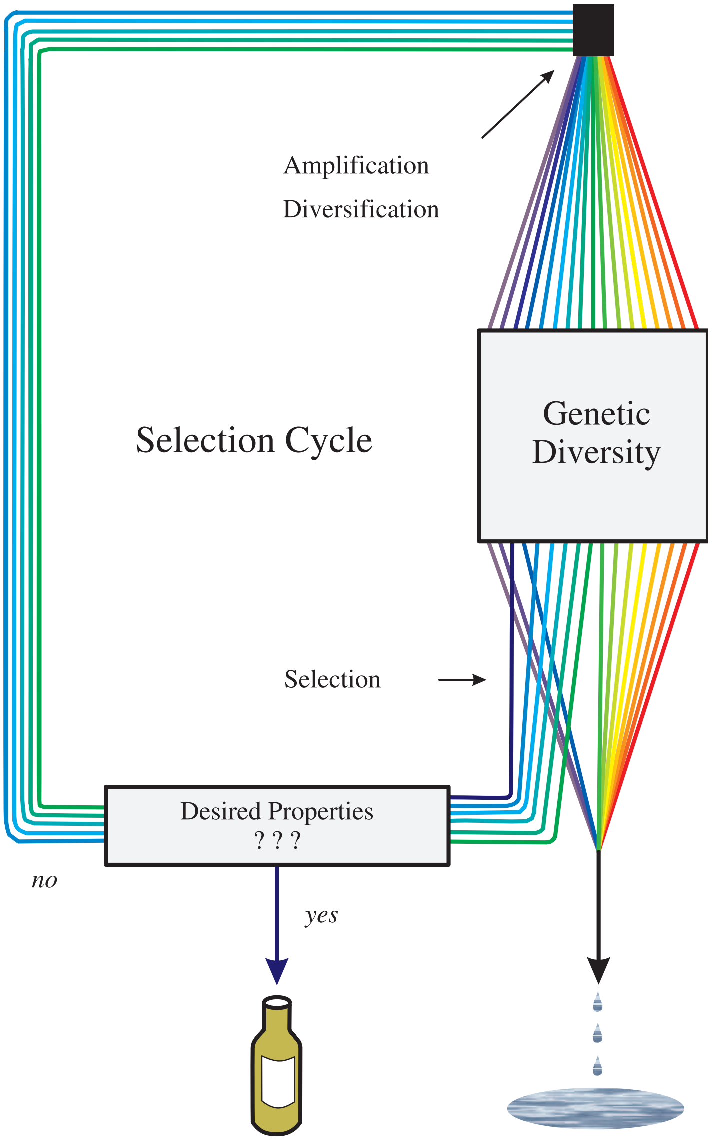
Since Spiegelman's pioneering works many different studies on evolutionary phenomena were performed with RNA molecules. RNA replication kinetics, for example, has been studied in detail. It is represented by a multistep polycondensation mechanism and the individual steps are now understood equally well as with other chemical reactions [5]. Serial transfer experiments in the presence of increasing concentration of the RNA degrading enzyme RNase A yielded "resistant" RNA molecules [67]. Catalytic RNA molecules of the group I intron family were trained to cleave DNA rather than RNA [4, 45], the SELEX technique [69] was applied to the selection of RNA molecules which bind to predefined targets with high specificity [36], and ribozymes with novel catalytic functions were derived from libraries of random RNA sequences [3, 7, 48]. Other approaches made use of *in vivo* selection to derive new variants of biopolymers [8, 49] or exploited the capacities of the immune system for evolutionary optimization in the design of catalytic antibodies [47, 54].

In fact, *in vitro* evolution experiments laid out the basis of a new discipline called evolutionary biotechnology (or applied molecular evolution) [14, 37, 38, 39]. In particular, RNA based evolutionary design of biomolecules has already become routine [18, 57]. In order to "breed" molecules for predefined purposes one cannot be satisfied with natural selection being tantamount to a search for the fastest replicating molecular species. In "artificial selection" the properties to be optimized must be disconnected from mere replication. The principle of evolutionary design of biomolecules is shown in figure 3. Molecular properties are optimized iteratively in selection cycles. The first cycle is initiated by a sample of random sequences or alternatively by a population derived from error prone replication of RNA (or DNA) sequences. Each selection cycle consists of three steps: (i) selection of suitable RNA molecules, (ii) amplification through replication, and (iii) diversification through mutation (with artificially increased error rates). The first step, selection of the genotypes that fulfil the predefined criteria best, requires biochemical and biophysical intuition or technological skill. Two strategies are commonly used: selection of best suited candidates from a large sample of different genotypes in homogeneous solutions or spatial separation of genotypes and massively parallel screening. Variants are tested and discarded in case they do not fulfil the predefined criteria. Selected genotypes are amplified, diversified by mutation with error rates adjusted to the current problem, and then the new population is subjected to selection. Optimally adapted genotypes are usually obtained after some twenty to fifty selection cycles. After isolation they can be processed by the conventional methods of molecular biology and genetic engineering.

4. RNA phenotypes and shape space covering

In *in vitro* evolution of RNA molecules genotypes and phenotypes are features of the same molecule, the sequence being the genotype and the spatial structure the phenotype. Mappings of genotypes into phenotypes are then reduced to the relation between sequences and structures of RNA molecules (fig.4). Current knowledge on three-dimensional structures of RNA molecules is rudimentary: only very few structures have been determined so far by crystallogaphy and nmr spectroscopy. Needless to say, spatial structures of RNA molecules are also very hard to predict. The so-called secondary structure of RNA is a coarse grained version of structure that lists only Watson-Crick and **GU** base pairs. It is conceptionally much simpler and allows to perform rigorous mathematical analysis and large scale computations. RNA secondary structure predictions are more reliable than those of full spatial structures. RNA secondary structures make it also possible to define formally consistent distance measures (η) in shape space [22, 34, 44, 51]. In addition, some statistical properties of RNA secondary structures were found to depend very little on choices of algorithms and parameter sets [68].

RNA secondary structures provide an excellent model system to study general relations between genotypes and phenotypes. The conventional approach of structural


Amplification
Diversification
Selection Cycle
Genetic Diversity
Selection
Desired Properties
? ? ?
no
yes