Landscapes
Complex Optimization Problems and Biopolymer Structures

by

PETER SCHUSTER[a,b,c,*] AND PETER F. STADLER[a,b]

[a]Institut für Theoretische Chemie, Universität Wien
[b]Santa Fe Institute
[c]Institut für Molekulare Biotechnologie, Jena

Abstract

The evolution of RNA molecules in replication assays, viroids and RNA viruses can be viewed as an adaptation process on a 'fitness' landscape. The dynamics of evolution is hence tightly linked to the structure of the underlying landscape. Global features of landscapes can be described by statistical measures like number of optima, lengths of walks, and correlation functions.
The evolution of a quasispecies on such landscapes exhibits three dynamical regimes depending on the replication fidelity: Above the "localization threshold" the population is centered around a (local) optimum. Between localization and "dispersion threshold" the population is still centered around a consensus sequence, which, however, changes in time. For very large mutation rates the population spreads in sequence space like a gas. The critical mutation rates separating the three domains depend strongly on characteristics properties of the fitness landscapes.
Statistical characteristics of RNA landscapes are accessible by mathematical analysis and computer calculations on the level of secondary structures: these RNA landscapes belong to the same class as well known optimization problems and simple spin glass models.
The notion of a landscape is extended to combinatory maps, thereby allowing for a direct statistical investigation of the sequence structure relationships of RNA at the level of secondary structures. Frequencies of structures are highly non-uniform: we find relatively few common and many rare ones, as expressed by a generalized form of Zipf's law. Using an algorithm for inverse folding we show that sequences sharing the same structure are distributed randomly over sequence space. Together with calculations of structure correlations and a survey of neutral mutations this provides convincing evidence that RNA landscapes are as simple as they could possibly be for evolutionary adaptation: Any desired secondary structure can be found close to an arbitrary initial sequence and at the same time almost all bases can be substituted sequentially without ever changing the shape of the molecule. Consequences of these results for evolutionary optimization, the early stages of life, and molecular biotechnology are discussed.

1. Evolutionary Adaptation

What are the minimal requirements for evolutionary adaptation? A very simple, if not the simplest, active reply to the environment is caused by the trial-and-error adaptation according to Darwin's principle of multiplication, variation and selection. The minimal requirements for Darwinian adaptation are easily stated:

- a population of objects that are capable of replication,
- a sufficiently large number of variants of those objects,
- occasional variations which are inheritable, and
- restricted proliferation which is constrained by limited resources.

According to our present day knowledge all inheritable variations are caused by changes of the genotype (i.e. by changes in the sequence of nucleotide bases of the DNA), and selection acts through differential reproductive success of phenotypes. Adaptive selection according to Charles Darwin requires nothing but non-vanishing differential fitness between variants. Variation, in essence, is based on two different principles: mutation and recombination. The first case represents a deviation in the genotype of the offspring from the parental genotype which is caused by erroneous replication. Recombination uses two parental genotypes and produces a mixed offspring by partial exchange of genetic information. Here, we shall be concerned only with variation through mutation.

The simplest entities that do actually fulfill all the four requirements listed above are RNA molecules. They are known to replicate *in vitro* in cell-free assays that contain activated monomers, a specific enzyme catalyzing replication (a replicase), and a suitable RNA molecule as template. The relevant properties of RNA molecules are outlined in section 2, a mathematical model for *in vitro* evolution is discussed in section 3. In section 4 we review the analogy between evolutionary optimization, combinatorial optimization, and statistical physics.

The efficiency of evolutionary adaptation, in essence, depends on the relation between genotypes and phenotypes. Sewall Wright[1] introduced the notion of a fitness landscape in order to illustrate evolution as an hill-climbing process on an (presumably) exceedingly complex surface. Recently the concept of fitness landscapes saw a revival in biology[2−5]. In section 5 we develop the theoretical framework for the study of (fitness) landscapes. The theory is then generalized to sequence-structure relation and applied to RNA folding in section 6. The implications of our results are discussed in section 7.

2. RNA Sequences and Structures

2.1. Template Chemistry

The capability for molecular replication is found with both classes of nucleic acid molecules, RNA and DNA. The principle is template copying by means of complementation of a single strand to a double helical structure. It is visualized best as an analogue of the conventional photographic (positive-negative) copying mechanism. Expressed in thermodynamic terms complementarity is the strong energetic preference for **G**≡**C** and **A**=**U** base pairs in the geometry of the double helix of nucleic acids that leads to an (almost) exclusive formation of these Watson-Crick-type base pairs. Boolean logic is very rare in chemistry and requires strong nonlinearities which have their origins in special structures (like here) or in special dynamics (for example in proof-reading mechanisms). Template chemistry is not restricted to nucleic acid molecules. Other examples of complementary template interactions were found recently. They represent candidates for non-natural molecular replication systems. Enzyme free, template induced RNA synthesis has been studied extensively by Leslie Orgel[6]. Template induced reproduction of RNA molecules occurs in cells that were infected by RNA viruses or viroids[7]. This is the simplest mechanism for molecular replication mechanism found in nature. Its biochemical details have been studied in great detail[8−10]. A single stranded RNA virus replicates in host cells using only a single enzyme. Double helix formation is merely an intermediate stage: the template and the newly synthesized strand are separated during the course of replication, and two RNA molecules, called *plus*- and *minus*-strand, are present after termination of the process.

2.2. Structure Formation

RNA molecules spontaneously form three-dimensional structures by folding the sequences in aqueous solutions, which contain appropriate concentrations of structure stabilizing divalent cations like Mg^{2+}, and have appropriate ionic strength, pH and temperature. The major driving force for structure formation is Watson-Crick base pairing (**G**≡**C**, and **A**=**U**) mediated by partial intramolecular complementarity of sequences, as well as **G**−**U** base pair formation. Other intermolecular forces and the interaction with the aqueous solvent shape the spatial structure of an RNA molecule. Small and medium size RNA molecules (with chain lengths $n < 200$) form equilibrium structures which are independent of the mechanism of folding and thus are completely determined by the sequence and the environmental conditions. Structure formation of large RNA molecules appears to be controlled at least in part by the kinetics of folding process.

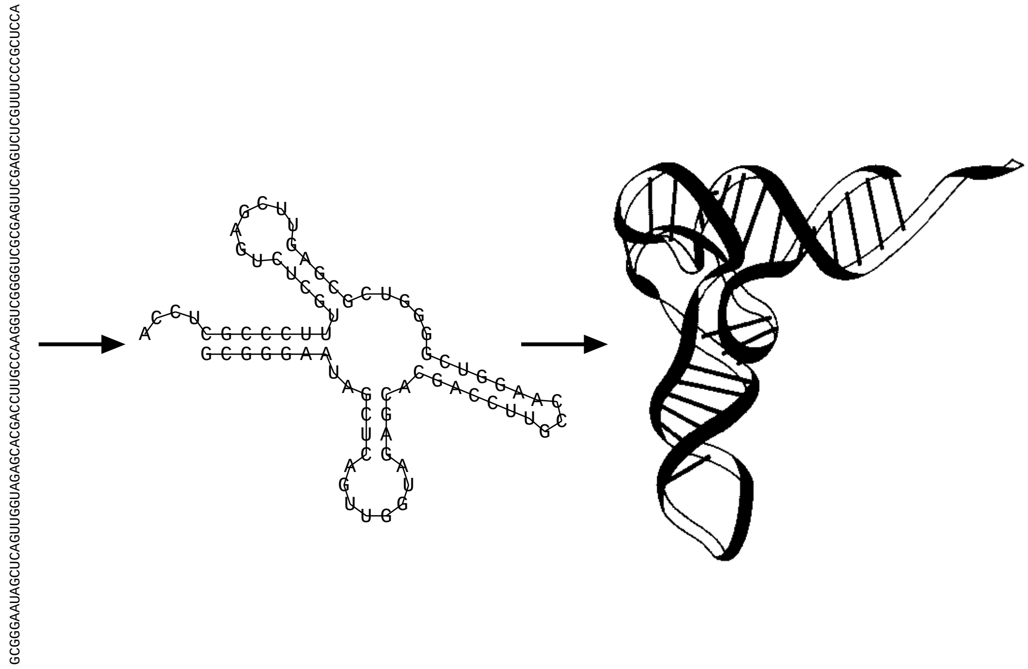

Figure 1: Folding of an RNA sequence into its spatial structure. The process is partitioned into two phases: in the first phase only the Watson-Crick-type base pairs are formed (which constitute the major fraction of the free energy), and in the second phase the actual spatial structure is built by folding the planar graph into a three-dimensional object. The example shown here is phenylalanyl-transfer-RNA (t-RNA $^{\mathrm{phe}}$) whose spatial structure is known from X-ray crystallography.

Formation of spatial structures of RNA molecules is commonly partitioned into two steps (Figure 1). The first step comprises conventional base pair formation (**AU**, **GC**, **GU**) and yields the so-called secondary structure of the RNA molecule, see section 2.3. In the second step the base pairing pattern is converted into the 3D structure.

2.3. Structure Prediction

Secondary structures are defined as listings of the Watson-Crick-type base pairs in the actual structure which fulfill the "no-knot condition" and may therefore be represented as planar graphs. Two features of RNA secondary structures will be important later on. They are discrete by definition (two bases either form pair or they don't), and they are composed of largely independent structural elements:

- *stacks*, i.e., double-helical regions,
- *loops*, i.e., unpaired regions enclosed by stacks (the degree of a loop is the number of stacks attached to it; consequently, a hairpin loop has degree 1, bulges and interior loops have degree 2, and all loops with degrees larger than 2 are multi-loops), and
- *external elements* are strains of unpaired bases that are not part of a loop (they are either joints, which connect components, or free ends).

There are several reasons for considering the secondary structure as a crude first approximation to the spatial structure of the RNA.

- Conventional base pairing and base pair stacking cover the major part of the free energy of folding.
- Secondary structures are used successfully in the interpretation of RNA function and reactivity.
- Secondary structures are conserved in evolutionary phylogeny.

The statistical investigation of RNA based landscapes requires the knowledge of several hundred thousand values derived from RNA structures. These data are not available at present, neither through experimental measurement nor through computation of the 3D-structures (which are highly time consuming and still unreliable). Restriction to secondary structures as a crude approximation to real structures, however, renders computation possible.

A variety of computer programs predicting RNA secondary structures have been published. A very brief overview is given in Table 1. Two public domain packages for RNA folding are currently available by anonymous `ftp`: Zuker's `mfold`[19], and the `Vienna RNA Package`[20], the latter providing a variety of different algorithms for structure prediction and structure comparison[21]. The data reported in this contribution have been produced by the `Vienna RNA Package`. See also section 4.3 for a discussion of the folding problem as a combinatorial optimization problem.

Table 1. Folding Algorithms.

Algorithm	ψ	Abbr.	Remark	Reference
deterministic				
Minimum Free Energy	$-$	`MFE`	fast	[11,12]
Kinetic Folding	$+$	`KIN`	fast	[13]
5'-3' Folding	$+$	`5-3`	fast	[14]
Partition Function	$-$	`PF`	ensemble	[15]
Maximum Matching	$-$	`MM`	unrealistic	[16]
stochastic				
Simulated Annealing	$+$	`SA`	very slow	[17,18]

$^{\psi}$ Pseudo-knots can be included. The major problem with the prediction of pseudo-knots is, however, the lack of sufficient experimental energy parameters.

2.4. Structure Comparison

RNA secondary structures can be represented as rooted planar trees [22−24]. The conversion assigns an internal node to each base pair and a leaf to each unpaired digit. An extra root is conveniently added in order to prevent the formation of a forest. An alternative tree representation (known as homeomorphically irreducible trees) is obtained by converting connected unpaired regions and connected stacked regions, respectively, into single nodes, that are given a weight corresponding to the size of the structural element. Both tree-types faithfully encode the secondary structure (for details see ref.[24]). Comparison of trees by means of edit distances is a well known technique in computer science[25,26].

An alternative method for the comparison of RNA secondary structures is based on the so-called mountain representation[27,28]. The secondary structures are encoded as linear strings with balanced parentheses representing the base pairs, and some other symbol coding for unpaired positions. Distances are computed by direct end-to-end alignment of these strings. Throughout this paper we will use $D(a,b)$ for a distance between secondary structures.

It has been shown that the results reported in this contribution are almost independent of the actual choice of the folding algorithm, the parameter set for the folding and the distance measures for secondary structures[29].

3. The Molecular Quasispecies

3.1. Mutation

In molecular evolution (and, apart from recombination, in all biology) the basis of variation is simply the limited accuracy of replication. Replication errors or mutations produce RNA sequences which differ from the parental template sequence. A change in the sequence will commonly cause a change in the structure. Mutations thus provide the genetic reservoir from which better adapted variants are chosen. They fall into three classes (Figure 2).

ACGAUGGGUUACC|G|AGGCAAGUCGUAG

Point mutation ⇓

ACGAUGGGUUACC|A|AGGCAAGUCGUAG

ACGAUG|GGUUACCG|AGGCAAGUCGUAG

Insertion ⇓

ACGAUG|GGUUACCG|GGUUACCG|AGGCAAGUCGUAG

ACGAUGGG|UUACCGAGGC|AAGUCGUAG

Deletion ⇓

ACGAUGGG|AAGUCGUAG

Figure 2: Three classes of mutations. Point mutations are copying errors with single base exchanges; they leave the chain lengths constant. In case of insertions part of the template sequence is duplicated during replication. A deletion leads to an error copy which is shorter than the original.

In reality the frequencies of all classes of mutations are phenotype dependent in the sense that the molecular structure of the RNA determines position dependent mutation rates. Position dependence of point mutations is much weaker than that of insertions and deletions, and therefore a uniform error model of point mutations was conceived and successfully applied to replication and mutation of RNA molecules *in vitro*, of viroids and of RNA viruses *in vivo*[30−33]. The model assumes that the error rate per (newly incoporated) base and replication event, p, is independent of the position in the sequence and of the nature of the point mutation (transition or transversion). It allows to express all mutation probabilities in terms of only three parameters, the copying fidelity $q = 1-p$, the mutant class h_{ki} (this is the number of positions in which the daughter sequence $\mathbf{I}_k$ differs from the parental sequence $\mathbf{I}_i$ which is commonly called the Hamming distance between the two sequences), and the chain length n:

$$Q_{ki} = p^{h_{ki}} q^{n-h_{ki}} = q^n \left(\frac{1-q}{q}\right)^{h_{ki}} . \tag{1}$$

No such model can be justified for insertions and deletions since they are directly related to the secondary structure of the RNA molecule. They occur very frequently at the beginning or at the end of hairpin loops. A plausible molecular interpretation postulates jumping of the replicase at these positions on the template in the forward direction (deletion) or in the backward direction (insertion).

3.2. Selection

The essentials of the kinetics of an *in vitro* evolution experiment can be captured in the reaction scheme described in figure 3. Eigen *et al.*[30−33] developed a deterministic theory under the constraint of constant total concentration of the macromolecular species $\mathbf{I}_k$:

$$\dot{x}_k = \sum_{j=1}^{N} Q_{kj} A_j x_j - D_k x_k - \overline{E} x_k , \qquad k = 1, 2, \ldots, N \tag{2}$$

where x_k is the relative concentration of the macromolecular species $\mathbf{I}_k$, A_k is its replication rate constant, and D_k is its degradation rate constant, respectively. The unspecific dilution flux, $\overline{E}$, which equals the net growth rate of the system, is given by

$$\overline{E} = \sum_{j=1}^{N} x_j (A_j - D_j). \tag{3}$$

$$
\begin{array}{rcl}
(A) + \mathbf{I}_k & \xrightarrow{Q_{kk} A_k} & 2\mathbf{I}_k \\
(A) + \mathbf{I}_k & \xrightarrow{Q_{lk} A_k} & \mathbf{I}_l + \mathbf{I}_k \\
\mathbf{I}_k & \xrightarrow{D_k} & \emptyset
\end{array}
$$

Figure 3: Reaction scheme for the quasispeces model. $\mathbf{I}_k$ denotes a macromolecular species. The energy-rich monomers *(A)* are assumed to be buffered and will therefore not appear explicitly in the rate equations. Their effect is included in the rate constants A_k and D_k.

The system of kinetic equations (1) has a unique globally stable equilibrium $\hat{x}$, whenever the mutation matrix is positive (or when the replication rates A_k are positive and Q is irreducible). Analytical solutions for the population structure $x(t)$ are known[34,35]. It is easily checked that $\overline{E}$ approaches the largest eigenvalue of the matrix W, defined by $w_{kj} = Q_{kj} A_j + \delta_{kj} D_j$, and $x(t)$ approaches a corresponding eigenvector. Eigen[30] defined the *fitness* of a macromolecular species $\mathbf{I}_k$ as w_{kk}.

3.3. Patterns in Sequence Space

Before we proceed with an analysis of the equilibrium population $\hat{x}$ we need to introduce a more geometric picture of the set of all sequences. The structure of the mutation matrix Q already suggests a quite natural "geometric" arrangement of the sequences: In fact, the h_{ik} provide us with an appealing measure for the distance between sequences. Restriction to point mutations (which commonly dominate in nature) implies that h_{ki} coincides with the Hamming distance[36], d_H, since it simply counts the number of positions in which two aligned sequences differ, i.e., it represents the minimum number of point mutations which are required to convert one sequence into another. The Hamming distance d_H is a metric for the abstract space of all sequences commonly called the *sequence space*. The concept of sequence space was first used in coding theory[37]. It was suggested for applications to proteins[38] and nucleic acids[39]. Later on it was expressed in quantitative terms

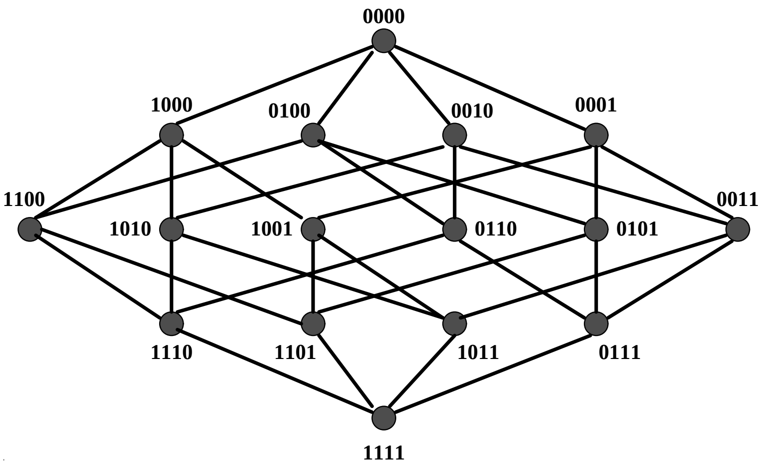

Figure 4: The sequence space of binary (**AU** or **GC**) sequences of chain length n=4. Every circle represents a single sequence of four letters. All pairs of sequences with Hamming distance d_H=1 (these are pairs of sequences that differ in only one position) are connected by a straight line. The geometric object obtained is a hypercube in four-dimensional space and hence, all positions (and all sequences) are equivalent. Furthemore all pairs of sequences with common mutual distance are geometrically equivalent, i.e., the error classes are symmetry classes of the hypercube graph.

for nucleic acids with different base pairing alphabets[31−33,40].

Sequence spaces are bizarre objects[40,41]: they are of very high dimension ($25 < n < 500$ for RNA molecules in test tube experiments, $250 < n < 400$ for viroids, and $3500 < n < 20\,000$ for (most) RNA viruses), but there are only few points on each coordinate axis. For the binary alphabet we have the hypercube of dimension n, see figure 4 for the example with $n = 4$. The distribution of sequences around an arbitrary reference sequence is described by

$$p(d) \;=\; \frac{(\kappa-1)^d}{\kappa^n}\binom{n}{h} \tag{4}$$

where κ is the number of letters in the alphabet and d is the index of the error class, i.e., the Hamming distance to the reference sequence. $p(d)$ increases exponentially for $d \ll n$, in sharp contrast to the quadratic or cubic increase on 2D or 3D lattices, respectively. We cannot expect therefore that our intuition, which is well trained with maps in three-dimensional space, will guide us well through these combinatorial spaces.

In the absence of replication errors we have selection for "fittest", i.e., for the species with the largest value of $A_k - D_k$[42,43]. Let us now consider the role of replication errors in population dynamics. Replication errors lead to new molecular species whose replication efficiency is evaluated by the selection mechanism. The

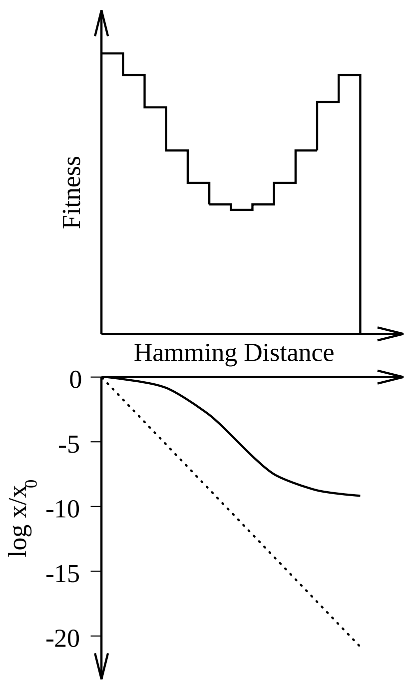

Figure 5: A single ridge in an otherwise flat fitness landscape has a tremendous effect on the population structure. In this example we have $n{=}50$ and a landscape that is flat except for 13 adjacent sequences with fitnesses shown in the upper half of the figure. For simplicity all degradation constants are set to 0. The equilibrium population structure is shown below (full line). The dotted line gives the mutant spectrum of the master species in the absence of the ridge for comparison.
In a typical laboratory population we expect to find even the 12-error mutant. If one of its neighbors were a new, fitter sequence it would be readily found and selected for. In the absence of a ridge the gap of Hamming distance 12 would effectively prohibit a transition, the better sequence in Hamming distance 13 is unlikely ever to be found in a lab experiment.

higher the error rate, the more mutations occur and the more viable mutants appear in the population. The stationary mutant distribution has been characterized as "quasispecies" since it represents the genetic reservoir of asexually replicating populations. The particular structure of a population depends crucially on the details of the landscape structure. For an example see figure 5.

For sufficiently small error rates p we have selection of the fittest, i.e., the sequence I_m with the maximum value w_{mm} will dominate the equilibrium population $\hat{x}$. The dominating species is commonly called *master sequence*. In general the master sequence is not necessarily the fittest sequence, since the selective advantage depends not only on its own fitness, but also on the influx terms $w_{mj}x_j$, which become larger when the mutant spectrum becomes broader. At not to small replication rates the population can be localized around a local optimum which has suboptimal fitness $w_l l$ by itself but which is surrounded by very fit neighbors. The transition between two master sequences is extremely sharp, see figure 6. This effect has been studied in detail by Schuster and Swetina[40,44].

The broardening of the mutant distribution makes evolutionary optimization faster and more efficient in the sense that populations are less likely caught in local fitness optima. There is, however, a critical error threshold[30−33,40]: if the error rate p exceeds this critical limit, heridity breaks down, populations are drifting in the sense that new RNA sequences are formed steadily, old ones dissapear, and no evolutionary optimization according to Darwin's principle is possible (Figure 6). There is an ordered (localized) population if

$$\sigma \cdot q^n > 1 \tag{5}$$

where the superiority of the master sequence, $\sigma \geq 1$, expresses differential fitness between the master sequence and the average of the remaining population. For the "golf course" landscape, defined by a single maximum (with fitness A_m) in an otherwise flat desert of fitness A_0, the superiority is simply $\sigma = A_m/A_0$. For more details on the relation of fitness landscape and superiority of the master species see section 5.6. In the limit $\sigma \to 1$ we are dealing with *neutral evolution*[45].

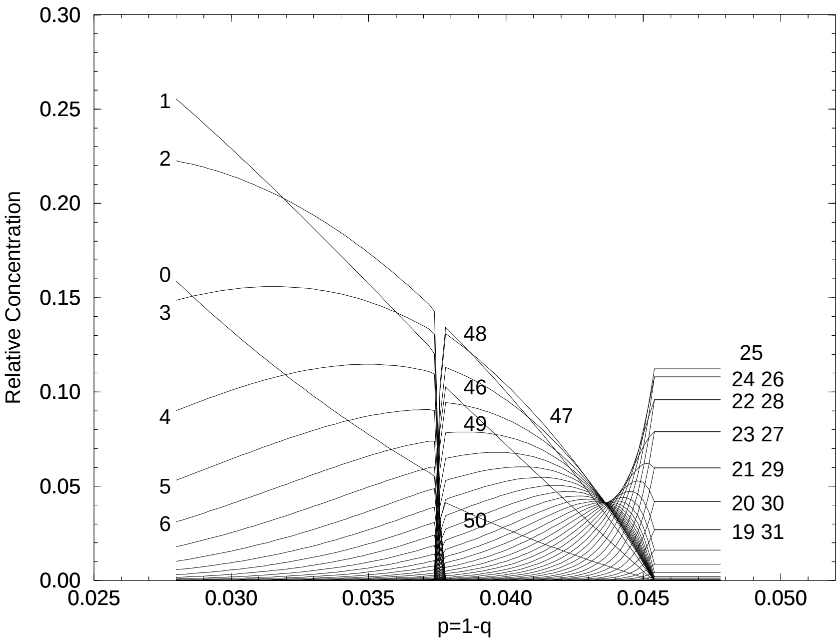

Figure 6: Cumulative concentrations of error-classes (labelled by the Hamming distance from the global optimum) in the equilibrium population for an almost degenerate quasispecies ($A_0=10, A_{50}=9.9, A_{49}=2, A_i=1$ otherwise). A transition between the master sequence $\mathbf{I}_0$ for low mutation rates, and $\mathbf{I}_{50}$ at intermediate error rates at $p_t \approx 0.0372$ is observed. The error threshold, indicated by the combinatorially dominating error class *25* becoming the most abundant in the population, lies at $p_{\rm cr} \approx 0.0454$. For $p \to 0$ the relative concentration of error class *0* approaches 1 (not shown).

The (evolutionary) adaptation of finite ensembles of sequences must be described by stochastic means for two reasons: (i) mutation is an intrinsically stochastic process, and (ii) the scarcely populated periphery of the mutant distribution forces a stochastic treatment there. It has been shown[46], however, that the results from the deterministic treatment outlined above carry over to the stochastic system with finite population size correction that vanish in the infinite population limit $M \to \infty$. The critical error rate, $r_{\rm cr}$, for instance, is given by

$$p_{\rm cr}(M) = p_{\rm cr}(\infty) - \mathcal{O}(M^{-1/2}). \tag{6}$$

Beyond the error threshold we obtain only time averages from the deterministic treatment. On a flat landscape there is a second critical value of the mutation rate, the *dispersion threshold*[47]. For moderate error rates we observe a fairly concertrated mutant cloud. The master sequence is no longer constant in time, however, but performes a random walk in sequence space. At high mutation rates beyond the dispersion threshold we have only random production of sequences, which fill the sequence space in a gas-like fashion. Numerical results indicate that the dispersion threshold also exists for models with complicated fitness functions[48,49]. The close relation between the quasispecies and spin glasses, and in particular between the error threshold and a phase transition has been investigated by Leuthäusser[50,51], Demetrius[52,53], and more recently by Tarazona[54].

3.4. Experimental Evidence

Selection experiments in molecular evolution were mainly carried out in well stirred homogeneous solutions. The material consumed by replication is renewed either discontinuously as in the serial transfer experiments[55] or continuously as in elaborate flow reactors[56,57]. Statistical removal of RNA molecules produced in excess restricts the populations size to be essentially constant (in the discontinuous serial transfer technique the population size is constant on the average).

An novel selection technique makes use of capillaries which contain a medium suitable for RNA replication. RNA is injected and a wave front spreads through the medium[58]. The front velocity of the traveling wave increases with the replication rate and hence, faster replicating mutants are selected by the wave propagation mechanism. The time axis of conventional evolution experiments is mapped onto a spatial axis by this technique.

The first experimental evidence that replicating ensembles of RNA molecules show adaptation of their structures to changes in the environment was provided by Sol Spiegelman and coworkers[59]. The medium for reproduction was deteriorated by the addition of a heterocyclic dye (ethidium bromide in the particular case) which intercalates between Watson-Crick base pairs, and thus interferes with template induced replication. Then the most frequently occurring genotypes (commonly called "master sequences") were isolated from populations and analyzed in the course of serial transfer experiments. A three error mutant of the original master sequence was found to replicate faster in the new medium. More recent investigations of replication in the presence of ribonucleases (these are enzymes which cleave RNA specifically in single stranded regions after certain bases) have shown directly how the structure of RNA molecules is adjusted in order to cope with detrimental changes in the environment[60]. Mutants appear and are enriched in the populations which compensate for the change in the environment by having fewer cleavage sites than the original master sequence or none at all.

Systematic kinetic studies on RNA replication and test tube evolution carried out by Christof Biebricher in Manfred Eigen's laboratory[8−10] revealed the mechanism of RNA replication. The rate and equilibrium constants which determine the outcome of selection under different conditions are known now. They can be determined independently from selection experiments by direct measurements. In addition minimum requirements for efficient replication were discovered. They consist in sequence regularities and sufficient secondary structure to facilitate double strand separation by the enzyme.

The 3D-structure of an RNA molecule determines its fitness in evolution experiments. The replicase recognizes the RNA template by its structure: only a subset of RNA structures binds to the enzyme sufficiently strongly for the initiation of replication. Small RNA species ($50 > n > 25$) that are replicated by the enzyme were isolated recently[61]. They share characteristic structural elements, a free 3'-end and a double helical region at the 5'-end of the RNA sequence and minimal requirements for the binding of RNA molecules to the replicase are derived from these data. All efficiently replicating RNA molecules have rich internal structure along the sequence which is apparently required to separate the plus- und the minus-strand during the process of replication. Speed and efficiency of several reaction steps of the replication process depend strongly on structural details of the enzyme and both RNA molecules, template and product. In general all rate and equilibrium constants of the replication process and hence also the over-all rate of RNA synthesis are functions of the 3D-structure. It determines therefore the fitness of an RNA molecule in molecular evolution experiments as the conventional phenotype does in organismic biology.

Adaptation, however, is not restricted to the optimization of molecular structures. Mutation rates can also be subjected to change. According to equation (5) such a change can be caused by a variation of the error rate p as well as by a change in the chain length n. Survival in variable environments demands sufficiently fast adaptation, and species with tunable error rates will adjust their quasispecies to meet the environmental challenge. In constant environments, on the other hand, such species will tune their error rates to the lowest possible values in order to maximize fitness.

Viruses are confronted with extremely fast changing environments since their hosts developed a variety of defense mechanisms ranging from the restriction enzymes of bacteria to the immune system of mammals and man. RNA viruses have been studied extensively. Their multiplication is executed by enzymes that do not allow large scale variations of replication accuracies ($q = 1 - p$). RNA viruses vary the error rate by changing the length of their genomes, and adjust the RNA chain lenghts to optimal values, which correspond to maximal chain lengths[30,31]

$$n_{\max} = \frac{\ln \sigma}{\ln q} \approx \frac{\ln \sigma}{p} . \tag{7}$$

Thus, RNA viruses adapt to their environments by driving optimization efficiency towards the maximum.

Adaptation in rapidly changing environments thus drives populations towards the error thresholds which is tantamount to the border between order (expressed in terms of quasispecies) and disorder in the form of drifting populations. Populations of RNA viruses were found to evolve indeed at the edge of disorder. Can we generalize this result to other organisms? In case Van Valen's "Red Queens hypothesis"[62,63] is met by actual ecosystems, coevolution of species would require each species to evolve at its maximum speed in order to compete successfully with evolutionary changes of the other species in the same ecosystem even under constant physical conditions, and the catchphrase "life is evolution at the edge of disorder" would be precisely at the point. We should keep in mind, however, that there are also "living fossils" in nature which have undergone hardly any macroscopically detectable change in hundreds of million years of life under apparently constant conditions. Under such constant environments optimization would favor small error rates since then the fitness of a given variant increases with increasing replication accuracy.

3.5. Recombination

Recalling RNA secondary structures and their importance for the replication process we recognize an important difficulty with recombination in evolution of RNA molecules. Unless the sites of recombination are directly related to the secondary structure of the template (for example, the sites have to lie in single stranded strechtes outside hairpin loops) the probability to produce viable descendants is very low. Recombination in RNA based genomes is practically unknown in nature except in cases where special mechanisms are available. The RNA virus *influenza A* is mentioned as a typical example: it has a split genome consisting of eight independent RNA molecules. Recombination occurs through simultaneous infection of a cell by two (or more) virus particles. Individual RNA molecules replicate independently and are randomly packed into the virus particles of the daughter generation. Genes from different parental viruses can be recombined in the virus progeny.

Although recombination is the basic ingredient of an important class of optimization algorithm, the *genetic algorithms*[64,65], we will restrict the present study to the case of mutation as the only source of variation.

4. Combinatorial Optimization

4.1. The Classical Optimization Problems

The travelling salesman problem (TSP)[66] is a classical example of an NP-complete[67] combinatorial optimization problem. Few mathematical problems have attracted as much attention. Given a distribution of cities the task is to find the shortest tour visiting each city once and returning to the starting point with prescribed costs η_{ij} for traveling from i to j. The cost function is

$$f(\tau) = \sum_{i=1}^{n-1} \eta_{\tau(i)\tau(i+1)} + \eta_{\tau(n)\tau(1)} \tag{8}$$

where τ is the permutation encoding the order of the cities. The symmetric problem $\eta_{ij} = \eta_{ji}$ has applications in X-ray crystallography[68], electronics[69] and the study of protein conformations[70]. The asymmetric case[71] has applications in scheduling chemical processes or from pattern allocation problems in glas industry.

The graph bipartioning problem is another well known NP-complete problem. It is much closer related to spin glass models than the TSP and its relatives. An instance is given by a graph with an even number n of vertices. A configuration is a partition of the vertices into two sets of equal size. The cost that has to be minimized is the number of edges connecting the two sets.

4.2. Spin Glasses

The class of most extensively studied spin glass systems consists of noble metall – transition metal alloys, e.g., *AuFe*, *AuMn*, *CuMn*, *AgMn*, etc. The concentration of the transition metal is typically between 0.1% und 10%. The nobel metal forms a matrix in which the transition metal occupies sites randomly. The energy of a particular configuration is typically modelled by a Heisenberg Hamiltonian of the form

$$\mathcal{H}(\sigma) = -\sum_{i,j} J_{ij}\, \sigma_i \cdot \sigma_j \tag{9}$$

where σ_i denotes the spin of the i-th transition metal atom. The disorder is "quenched", i.e., the coupling constants remain fixed for a given sample. A large variety of models differing by the choice of the J_{ij} have been studied, see, e.g., refs. [72,73]

The Heisenberg Hamiltonian can be generalized to include the interaction of p spins at a time. The Hamiltonian has the form

$$\mathcal{H}(\sigma) = \sum_{i_1 i_2 \dots i_p} J_{i_1 i_2 \dots i_p} \sigma_{i_1} \sigma_{i_2} \dots \sigma_{i_p} \tag{10}$$

In a certain sense the limit $p \to n$ corresponds to Derrida's[74] random energy model, in which the energy of a spin configuration (microstate) is an i.i.d. Gaussian random number. There is a large class of disordered solids related to spin glasses, such as electric dipolar and quadrupolar glasses.

Kauffman's Nk model[75,76] is closely related to spin glass Hamiltonians. A configuration is given by a binary string b of length n. Each position i is assigned k "neighbors" which influence site i. For each choice of the bits at position i and its neigbors a random site fitness f_i is drawn independently from some distribution. The fitness of the configuration b is defined as the average over the site-fitnesses. The major adavantage of the Nk model is that the ruggedness of the fitness function can be tuned by k.

4.3. Folding of Biopolymers

Linear biopolymers fold into 3D structures which depend on the particular sequence of the monomers. The folding is commonly viewed as an optimization problem for an energy function which is composed of a large number of local energy contributions. We remark, however, that it is still unclear whether the structure of large polymers is determined by the free energy, or whether it is defined by the kinetics of the folding process. In the case of RNA *secondary* structures the energy minimizations can be done using a dynamics programming scheme, see 2.3; no reliable algorithm exists, however, for the prediction of 3D structures. The case of proteins is even worse: polynomial time optimization algortithms are unknown, even for the simplest lattice models[77−79]. The energy functions themselves, however, can be characterized statistically. Such studies are currently pursued in our group.

In the next section we shall develop a theory that allows to compare and classify models of the above types. We will provide some measures that enable us to decide whether or not a model landscape is a reasonable approximation to a biological fitness landscape.

5. Towards a Theory of Combinatory Landscapes

5.1. Definition

Two features are common to all landscapes:

(1) A *cost function* or *fitness function* f assigns a value to each *configuration*,

$$f : \mathcal{C} \to \mathbb{R}, \tag{11}$$

e.g, an energy to each spin configuration, a fitness value to each RNA sequence, a length to each tour in a traveling salesman problem.

(2) There is a rule that determines whether two configurations x and y are nearest neighbors.

A list of corresponding terms for landscapes occurring in different fields are is presented in table 2.

Table 2. The Correspondence of Statistical Mechanics, Combinatorial Optimization and Evolutionary adaptation.

Combinatorial Optimization	Statistical Mechanics	Folding of Biopolymers	Evolutionary Adaptation	Symbol
system size	system size	chain length	chain length	n
configuration	micro-state	structure	sequence	x
cost	energy	energy	fitness	$f(x)$
pseudo-temperature	temperature	temperature	mutation-rate	$T = -\ln q$
optimum	ground state	native state	fittest	o
instance	sample	*	*	f
move	(spin flip)†	†	mutation	

† There is no generic term for the elementary process of the dynamics
* Ideally, there is only one energy or fitness function in a given system

As a consequence of (2), the set $\mathcal{C}$ of all configurations can be viewed as a graph with each vertex corresponding to a configuration and an edge between any two nearest neighbors. Therefore, there is a natural metric distance $d(x, y)$ between any two configurations given by the shortest path in the graph $\mathcal{C}$. This is a straight forward generalization of the construction of the sequence spaces above.

Two vertices x and y are *equivalent* if there is a graph a automorphism, i.e., a symmetry operation, α, such that $\alpha(x) = y$. (Since the automorphisms form a group we have also $y = \alpha^{-1}(x)$.) In this paper we will exclusively deal with configuration spaces where all vertices are equivalent. Such graphs are called vertex transitive.

Two pairs of vertices (x, y) and (u, v) are equivalent if there is a graph automorphism such that $x = \alpha(u)$ and $y = \alpha(v)$. If the graph is vertex transitive we may choose an arbitrary vertex as reference point. The set of vertices then decomposes into symmetry classes in complete analogy to the error-classes in sequence spaces. The set of all pairs of vertices can hence be partioned into M symmetry classes. The number M is at least the maximum distance in the graph plus 1 (in the case of distance transitive graphs) and at most the number of vertices (in case that no two pairs containing one common vertex are equivalent). These symmetry classes play a key role in the theoretical investigation of landscapes.

The model landscapes presented in the previous section contain a stochastic element. A particular instance is generated by assigning a usually large number of parameters at random. These landscapes are sample functions from a random field $\mathcal{F}$ which is defined by a probability distribution $P(c_1, c_2, \ldots, c_N)$ that $f(x_i) < c_i$ for all $i = 1, \ldots N$. Consequently we can in principle calculate the expected value of any property X by

$$\mathcal{E}[X] = \int_{\mathbb{R}^N} X \, \mathrm{d}P(c_1, \ldots, c_N). \tag{11}$$

The alternative point of view is to stick to a particular instance of a landscape and persue a statistical analysis of this instance. This is the only possibility when a suitable ensemble is unknown or does not exist. This is the case, e.g., for the landscapes of protein folding, viral adaptation, and for some technical optimization problems such as the "low autocorrelated string problem". In this situation we have to be content with averages over values taken from our single instance. For example we have

$$\mathrm{Var}[f(x)] = \mathcal{E}[f(x)^2] - \mathcal{E}[f(x)]^2 \qquad \text{and} \qquad \sigma^2 = \langle f^2 \rangle - \langle f \rangle^2, \tag{12}$$

where the first average is taken over all possible instance of f at a fixed position x, while the second average runs over all vertices x in a fixed instance f. Of course the two variances $\mathrm{Var}[f(x)]$ and σ^2 are not the same in general. If, for a given quantity X, both averages give the same result, one says that X is *self-averaging*[80].

So far all investigations of landscapes were resticted to the first two moments for two reasons: (1) most model landscapes have a normal or nearly normal distribution and hence no information can be gained from considering higher moments, and (2) the computational problems for numerically investigating higher moments are prohibitiv. The following subsection will therefore deal exclusively with the "second order" properties of landscapes.

5.2. Isotropy

The role of ridges and plateaus for the velocity of adaptation is well known in the literature[32,33]. See also section 3.3 for an example. Such special structures form outstanding landmarks in a landscape. They are not likely to occur in an "isotropic" model landscape. In this section we will make precise the notion of an "isotropic landscape".

Definition. A landscape is *isotropic* if the following two conditions hold.
(i) $\mathcal{E}[f(x)] = \bar{f}$ for all configurations $x \in \Gamma$.
(ii) For any two equivalent pairs of configurations (x, y) and (u, v) holds $\mathcal{E}[f(x)f(y)] = \mathcal{E}[f(u)f(v)]$.

Weinberger[81] suggested a more restrictive definition of isotropy requiring that instead of (ii) holds:
(ii') for any two pairs of configurations (x, y) and (u, v) with common distance, $d(x, y) = d(u, v)$, holds $\mathcal{E}[f(x)f(y)] = \mathcal{E}[f(u)f(v)]$.

We will call a landscape fulfilling (ii') *homogeneous*. Note that on distance transitive configuration spaces, such as the sequences space, both definitions are equivalent. In general the two notions are not equivalent.

There are three reasons for using the less restrictive definition:
1) It seems natural to assume that the correlation of two pairs of vertices is not the same if they are not equivalent in configuration space.
2) The formalism for isotropic landscapes works as well when only (ii) is assumed.
3) As seen in the above example the landscapes of simple models on more complicated graphs are still termed isotropic. In particular, the condition for obtaining an AR(1) landscape on a non-distance-transitive graph implies in general different covariances for pairs of vertices in different symmetry classes which may well belong to the same distance class. Condition (ii') would then prohibit the existence of isotropic AR(1) landscapes (see below). For instance,

the landscape of a symmetric TSP with inversions is isotropic but not homogeneous.

Examples of isotropic landscapes are long range spin glasses, in particular the Sherrington-Kirkpatrick model, the graph bipartioning problem, and most variants of Kauffman's Nk-model.

Example. Let us consider an Nk model with random assignment of neighbors (see section 3.2). As a simple consequence of the definition of the Nk model we find that both $\mathcal{E}[f(x)]$ and $\mathcal{E}[f^2(x)]$ are independent of the configuration x. Let b be an arbitrary bit string and let b^l be the mutant obtained by flipping bit number l. Equation (17) in the next section shows than the landscape cannot be isotropic if $\Delta_l^2 = \mathcal{E}[(f(b) - f(b^l))^2]$ depends on l. It is easy to check that $\Delta_l^2 = 2s^2/n^2 \cdot \mathcal{E}[|U(l)|]$ where $U(l)$ denotes the set of sites which are influenced by flipping the l-th bit. When the neighborhoods are assigned in such a way that the k positions contributing to a given site's fitness are picked at random, then we cannot expect that each bit will be chosen the same number of times. If the same assigment of neighbors is used for all instances of the landscape, the model is not isotropic. If, on the other hand, the neighbors are independently chosen for each instance then $\mathcal{E}[|U(l)|] = k + 1$ (k neighbors plus position l itself) and Δ_l^2 is independent of l, i.e., the model is isotropic.

5.3. Correlation in Isotropic Landscapes

Correlation functions have proven to be very useful for the classification and analysis of fitness landscapes. They have been studied for a large number of model landscapes taken from spin glasses and combinatorial optimization problems (Table 3), and for various RNA landscapes (Table 4). In accordance with previous investigations[82] we use the following definition for the autocorrelation function of an isotropic landscape $(\mathcal{C}, f)$ given by

$$\begin{aligned} \rho(\mu) &= \frac{\mathcal{E}[f(x)f(y)] - \mathcal{E}[f(x)]\mathcal{E}[f(y)]}{\mathrm{Var}[f]} \quad \text{with} \quad (x,y) \in \mu \\ \hat{\rho}(\mu) &= \frac{\langle f(x)f(y)\rangle_\mu - \langle f\rangle_\mu^2}{\sigma^2} \end{aligned} \tag{13}$$

where μ is a symmetry class of pairs of configurations. The averages $\langle\,.\,\rangle_\mu$ run over all pairs in the partition μ. Note that for distance transitive graphs, like sequence spaces, the partition μ is simply a distance class.

For an instance of an isotropic landscape holds $\langle f(x)f(y)\rangle_\mu = \mathcal{E}[f(u)f(v)]$ for all $(u,v) \in \mu$. Let $p(\mu)$ be the probability for picking a pair of vertices in symmetry class μ at random. Then

$$\overline{\rho} = \sum_\mu p(\mu)\rho(\mu) \tag{14}$$

is the back ground correlation in the landscape. This quantity does not necessarily vanish! It is easy to show, however, that the empirical autocorrelation function $\hat{\rho}$ always fulfills

$$\sum_\mu p(\mu)\hat{\rho}(\mu) = 0 \tag{15}$$

As a consequence one can show that the empirical variance and the autocorrelation function of the landscape are related to the the ensemble variance and the ensemble autocorrelation function, respectively:

$$\sigma^2 = \mathrm{Var}[f]\,(1-\overline{\rho}) \qquad\qquad \hat{\rho}(\mu) = \frac{\rho(\mu)-\overline{\rho}}{1-\overline{\rho}} \tag{16}$$

Therefore we have the following

Theorem. Let $(\mathcal{C}, f)$ be isotropic. Then the following statements are equivalent
(i) The back-ground correlation $\overline{\rho}$ vanishes
(ii) The variance is self-averaging, $\sigma^2 = \mathrm{Var}[f]$.
(iii) The autocorrelation function is self-averaging, $\hat{\rho} = \rho$.

An alternative representation of the autocorrelation function has been very useful in previous studies

$$\rho(\mu) = 1 - \frac{\mathcal{E}[\,(f(x)-f(y))^2\,]}{2\mathrm{Var}[f]} \qquad \hat{\rho}(\mu) = 1 - \frac{\big\langle \big(f(x)-f(y)\big)^2 \big\rangle_\mu}{2\sigma^2} \tag{17}$$

We use these expressions below to calculate the exact autocorrelation function for travelling salesman problems. They have been used successfully to generalize the notion of an autocorrelation function to the mapping from sequence space to (RNA) secondary structures, see section 6.3.

Example. A common representation of the TSP uses a string of lenght $n(n-1)/2$ to code for the tours of a symmetric TSP. For an asymmetric TSP we need a string of length $n(n-1)$ because the bonds $(i \to j)$ and $(j \to i)$ are distinct in this case. Each position in this string corresponds to a bond, i.e., to a pair of cities. Let

Table 3. Combinatorial optimization problems and their autocorrelation functions.

Problem	Reference	Metric	diamΓ	ℓ	$r(s)$	$\rho(d)$	Reference
Random Energy	74	(any)	–	0	%	$\delta_{0,d}$	-
symmetric TSP	66	Transpositions	$n-1$	$n/4$	$\sim e^{-4s/n}$		83
		Reversals	$n-1$	$n/2$	$\sim e^{-2s/n}$	$1-d_H(x,y)/\overline{d_H}$	83
		Canonical Tr.	$\frac{n(n-1)}{2}$	$n/2$	$\sim e^{-2s/n}$		84
asymmetric TSP	66	Transpositions	$n-1$	$n/4$	$\sim e^{-4s/n}$		83
		Reversals	$n-1$	–	$\sim\frac{1}{2}(\delta_{0,s}+e^{-2s/n})$	$1-d_H(x,y)/\overline{d_H}$	83
		Canonical Tr.	$\frac{n(n-1)}{2}$	$n/3$	$\sim e^{-3s/n}$		84
Graph Matching	85	Transpositions	$n-1$	$n/4$	$\sim e^{-4s/n}$	?	86
Graph Bipartitioning	87	Exchanges	$n/2$	$(n-3)/8$	$(1-\frac{8}{n}+\frac{8}{n^2})^s$	$1-\frac{n-1}{n-2}[8\frac{d}{n}-16(\frac{d}{n})^2]$	88
Low Autocorrelated Binary Strings	89	Hamming	n	$\approx n/10$	$\sim e^{-10s/n}$	?	87
adjacent Nk	75 76	Hamming	n	$n/(k+1)$	$\sim(1-\frac{k+1}{n})^s$		90
random Nk	75 76	Hamming	n	$n/(k+1)$	$\sim(1-\frac{k+1}{n})^s$	$(1-\frac{d}{n})(1-\frac{k}{n-1})^d$	90
pure random Nk	91	Hamming	n	$n/(k+1)$	$\sim(1-\frac{k+1}{n})^s$	$1-(\frac{k+1}{n})^d$	91
p-Spin	92	Hamming	n	$n/(2p)$	$(1-\frac{2p}{n})^s$	$\frac{1}{\binom{n}{p}}\sum_j(-1)^j\binom{d}{j}\binom{n-d}{p-j}$	85
							93
SK-glass	94	Hamming	n	$n/4$	$(1-\frac{4}{n})^s$	$1-\frac{n}{n-1}[4\frac{d}{n}-4(\frac{d}{n})^2]$	85

% The autocorrelation function along a random walk depends on the geometry of the configuration space.

us define the indicator function χ by $\chi(a,x) = 1$ if bond a occurs in tour x and $\chi(a,x) = 0$ otherwise. We find

$$\mathcal{E}[(f(x) - f(y))^2] = \sum_a \mathcal{E}[L_a^2](\chi(a,x) - \chi(a,y))^2 \tag{18}$$

with $\chi(a,x) = 1$ if bond a occurs in tour x and $\chi(a,x) = 0$ otherwise. L_a is the length of bond a. The term in brackets is 1 iff bond a occurs in exactly one of the tours, i.e., we have as many nonzero term as there are differing positions in the string representation. Therefore $\mathcal{E}[(f(x) - f(y))^2] = \mathcal{E}[L^2] \cdot d_H(x,y)$ where d_H denotes the Hamming distance of the string representations of the tours. The average Hamming distance over all pairs of configurations,

$$\overline{d_H} = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} d_H(x,y) \tag{19}$$

is of course independent of x. The autocorrelation in any TSP landscape is therefore given by

$$\rho_{TSP}(x,y) = 1 - \frac{d_H(x,y)}{\overline{d_H}} \tag{20}$$

Note that this relation does not depend on the move set. The correlation of the landscapes induced by different move sets depend therefore on the the relation of the edit-distance and the Hamming distance of pairs of tours. This is easy to see for the nearest neighbor correlations: Since we have $\overline{d_H} \approx 2n$, and since almost all pairs of tours do not share common bonds and the exchange of one bond by another one causes a Hamming distance of 2, we find that the nearest neighbor correlation is given by $\rho_{TSP}(1) \approx 1 - B/n$, where B is the number of bonds exchanged by a single move[84].

In order to compare landscapes it is desirable to condense the information of the autocorrelation function into a single number. Most landscapes show smoothly decaying autocorrelation functions which (for not too large distances) can be approximated by a decaying exponential. We *define* the characteristic length ℓ of a landscape by

$$\rho(\ell) = 1/\mathrm{e}. \tag{21}$$

An alternative measure would be the initial slope of the autocorrelation function, $\rho' = 1/\lambda$. For most landscape we find empirically $\ell \approx \lambda$.

Table 3 shows that for most landscapes the correlation length scales linearly with the systems size, $\ell = n\ell^* + c$. It is often useful to scale the distance by the diameter of the configuration space, $\xi = d/\mathrm{diam}\, \mathcal{C}$. The autocorrelation function of the scaled variable will be denoted by $\rho^*(\xi) = \rho(d)$. The scales characteristic length is $\ell^* = \ell/\mathrm{diam}\, \mathcal{C}$.

Example. Various kinds of RNA landscapes have been investigated in detail[95,96]. Scaled autocorrelation functions for the **GCAU** and the **GC** landscapes are shown in figure 7.

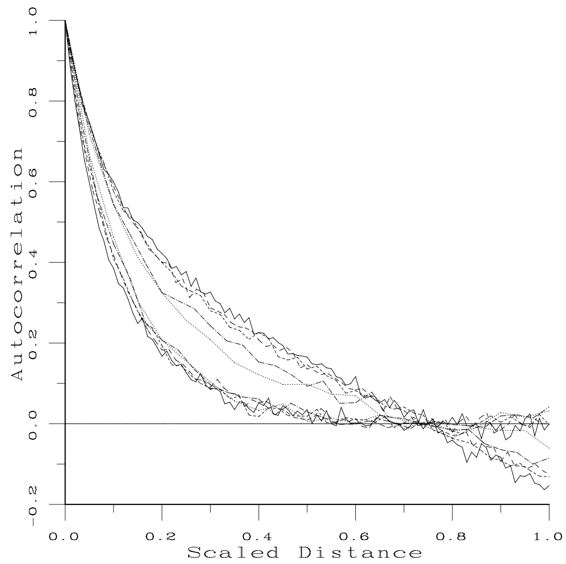

Figure 7: Empirical autocorrelation function for **GCAU** (upper set) and **GCXK** (lower set) free energy landscapes as function of the scaled Hamming distance $\xi = d/n$. The line type indicates the chain length: $n=20$ *dotted*, $n=30$ *dash-dotted*, $n=50$ *short dashed*, $n=70$ *long dashed*, $n=100$ *solid*. The data indicate that the autocorrelation function converge towards a limit function that is characteristic for the type of the landscape.

Correlation lengths of free energies are essentially linear functions of the chain length n. Data for the scale of correlation lengths are compiled in table 5. The base pairing alphabet has remarkably strong influence on the correlation length: the ℓ-values for **AUGC**-sequences are much longer than the corresponding values for pure **GC**- or pure **AU**-sequences. The optimization of properties of RNA molecules derived from the natural **AUGC** alphabet is therefore much easier than that of RNA molecules predominantly consisting of **G** and **C** or **A** and **U**, respectively. The dominant contribution apparently comes from the base pairing strength and not from the logic of the pairing, as can be seen from the large difference of the **GCAU** and the **GCXK** alphabet.

Table 4. Scaled correlation lenght ℓ^* for RNA free energy landscapes.

	ΔG				$\Delta G^\sharp$		Model†	
Alphabet	MFE	MFE[s]$^+$	PF	KIN	Melt.	Rec.	D	A
GC	0.0857	0.0768	0.127	0.0448	0.099	0.030	$\leq 3.5^*$	$\leq 3^*$
AU	0.0600							
GU	0.0514							
AUI	0.0532							
AGC	0.0982							
AUC	0.0856							
AUG	0.1191							
UGC	0.3850							
GCAU	0.2627	0.2976	0.243	0.1734	0.268	0.055	0.2	$\leq 15^*$
GCAU$^\%$	0.2044							
GCXK	0.1070				0.125	0.070		
ABCDEF	0.1182							

All data refer to T=37°C.

$^+$ [s] refers to Salser's old parameter set[97], all other data have been obtained with a recently updated paramter set[98].

* A linear dependence on the chain length could not be verified. The bounds correspond to chainlenght $n \leq 120$.

† These landscapes have been used to model replication and degradation in the computer simulations by Fontana *et al.*[99,100]

$^\%$ **GU** pairs forbidden.

Weinberger[101] proposed using the autocorrelation function $r(s)$ of values sampled along a simple random walk $x_0, x_1, x_2, \ldots$ on the configuration space $\mathcal{C}$ (A simple random walk means that each neighbor is chosen with the same probability). Let $r(s)$ be the autocorrelation function of the time series $\{f(x_0),\ f(x_1),\ f(x_2),\ \ldots\}$. On vertex transitive graphs the functions $\rho(\mu)$ and $r(s)$ are closely related via the geometric relaxation of the random walk on the configuration space. Let $\varphi_{s\mu}$ denote the probability that the pair (x_0, x_s) belongs to the symmetry class μ. Then we have

$$r(s) = \sum_\mu \varphi_{s\mu}\rho(\mu). \tag{22}$$

Both, numerical experiments and analytical calculations show that the correlation functions $r(s)$ for model landscapes are (at least very close to) decaying exponentials. Weinberger[101] coined the term AR(1) landscape for landscapes with Gaussian distribution of values and an autocorrelation function of the form a^s. It has been shown recently that an autocorrelation function of this form is a symmetry property of landscape and that only selected values of the basis a can occur for isotropic landscapes. In fact, a landscape shows an autocorrelation function $r(s)$ of

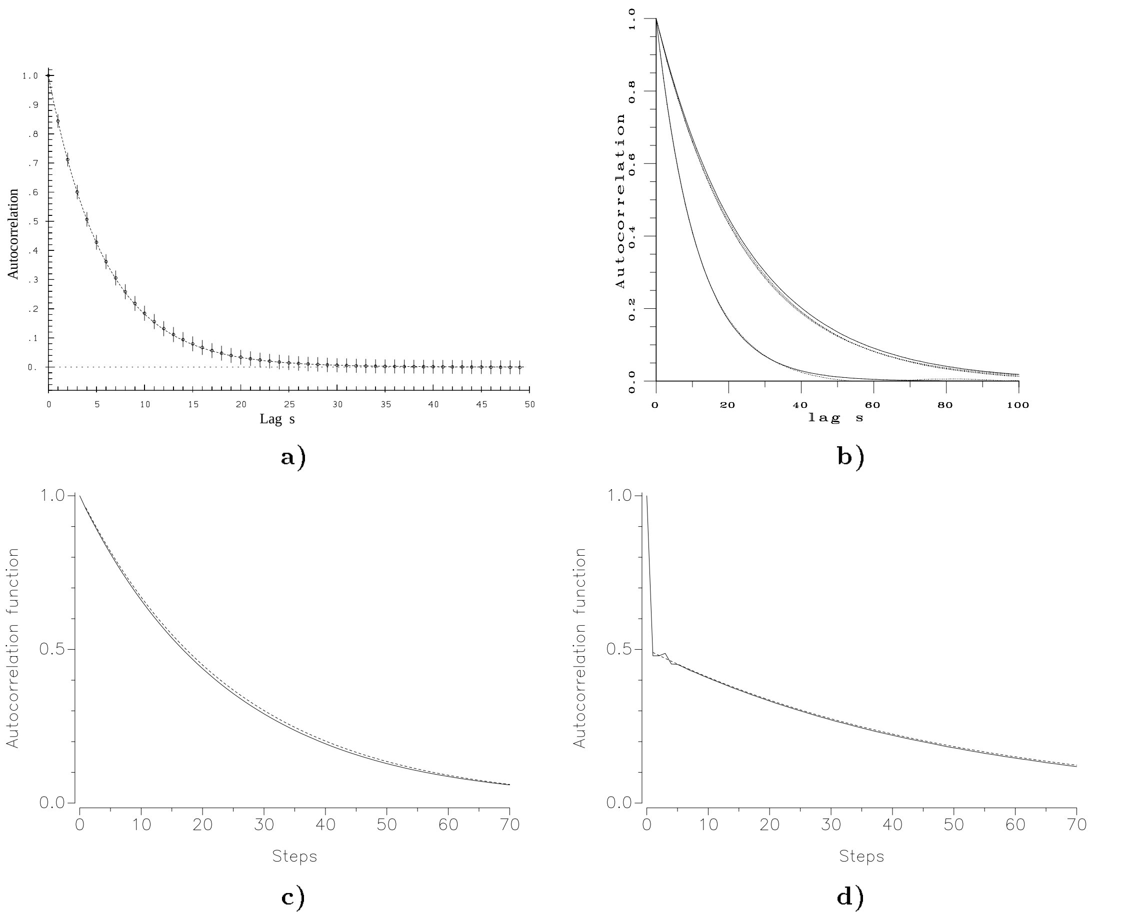

Figure 8: Some examples of autocorrelation functions for model landscapes. a) GBP, b) graph matching (upper curves) and low autocorrelated binary string problem, c) TSP with transposition metric, d) asymmetric TSP with reversal metric.

this form if and only if $\rho(\mu)$ is an eigenvector of the so-called "collapsed adjacency matrix" of the configuration space. In this case the basis a is the corresponding eigenvector of this matrix devided by the number of nearest neighbors (for details see refs.[93,102]).

5.4. Empirical Anisotropy

A serious problem arises when we do not know the ensemble from which our example is drawn. It is crucial for the understanding of the optimization dynamics in such a landscape to know whether it looks like a typical instance of an isotropic random field or whether there are significant anisotropies.

Recently a method for determing the degree of anisotropy has been suggested[103]. In order to study "local" features of the cost function f on $\mathcal{C}$ a system $\mathcal{B}$ of test sets with the following properties is introduced:
(1) $\mathcal{B}$ is a partition of the configuration space.
(2) $A \in \mathcal{B}$ is a connected subgraph of $\mathcal{C}$.
(3) Any two subgraphs $R, S \in \mathcal{B}$ are isomorphic.

These quite restrictive requirements are imposed to exclude any bias resulting from the different geometries of the test sets. Furthermore we require that the test sets are small compared to the full configuration space but still contain a large number of configurations: $1 \ll |A| \ll |\mathcal{C}|$. The average correlation of pairs of configurations in a test set, $\bar{\rho}_{\mathcal{B}}$, will not vanish in general. This reflects that the connected sets A are much smaller than the entire configuration space $\mathcal{C}$.

Definition. A value landscape is *empirically isotropic* if for any family $\mathcal{B}$ of test sets holds

$$\left\langle \big(f(x) - f(y)\big)^2 \right\rangle_\mu^A = \left\langle \big(f(x) - f(y)\big)^2 \right\rangle_\mu, \tag{23}$$

i.e., the average of the squared fitness difference of pairs of configurations belonging to a given symmetry class μ will be the same whether it is measured in single test set or over the entire configuration space. This condition can be replaced by $\mathrm{var}_{\mathcal{B}}[\langle f \rangle_A] = \sigma^2 \cdot \bar{\rho}_{\mathcal{B}}$, where $\mathrm{var}_{\mathcal{B}}[\langle f \rangle_A]$ is the variance of the mean fitnesses of the test sets and σ^2 is the variance of the landscape. It seems natural to measure the anisotropy of a landscape by the extent to which this relation is violated. We therefore call the dimensionless parameter

$$\alpha = \frac{\mathrm{var}_{\mathcal{B}}[\langle f \rangle_A]}{\sigma^2} - \bar{\rho}_{\mathcal{B}} \tag{24}$$

the *coefficient of anisotropy* (with respect to the partition $\mathcal{B}$).

Example. Consider a Sherrington-Kirkpatrik spin glass with $2n$ spins. The landscape of the Sherrington-Kirkpatrick glass is known to be isotropic. The autocorrelation function ρ is given in table 1. A reasonable partition $\mathcal{B}$ is obtained by

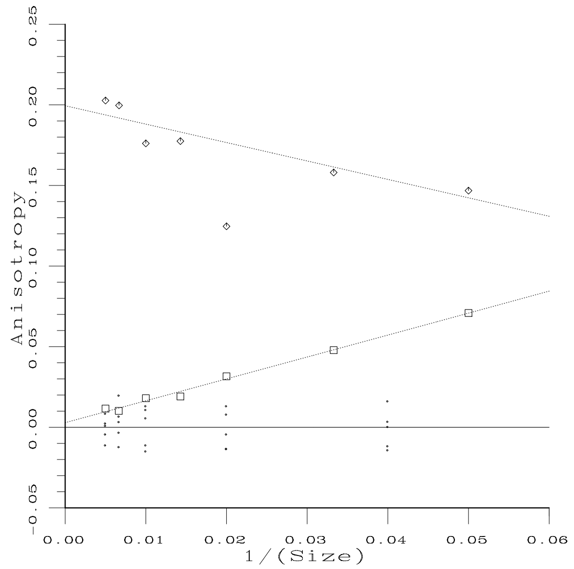

Figure 9: Empirical coefficient of anisotropy as a function of chain length.
Symbols: ◇ **GCAU**, □ **GCXK**, · Sherrington Kirkpatrick spin glass.

fixing half of the spins within a given set $A \in \mathcal{B}$. The results of numerical experiments in figure 9 show that the instances of the SK model are indeed empirically isotropic.

Example. Consider the sequence space consisting of all sequences of length n, which can be built on a four letter alphabet. We compare the RNA free energy landscape of the natural alphabet $\{$**G**, **C**, **A**, **U**$\}$ with the landscape of the artificial alphabet $\{$**G**, **C**, **X**, **K**$\}$. The bases **X** (Xanthine) and **K** (2,6-diaminopyrimidine) for a base pair with about the same characteristics as the **G≡C** pair[104]. We construct the test sets as follow: A_σ is the set of all sequences that have **G** or **C** at each position i with $\sigma_i = 1$ and **A** or **U** (or **X** or **K**, respectively) on the remaining positions ($\sigma_i = 0$). The binary string σ therefore characterizes the "slice" A_σ which itself is a Boolean Hypercube with dimension n. Numerical data for the coefficient of anisotropy are shown in Figure 9. While the artifical alphabet gives rise to an asymptotically isotropic landscape, we find a large amount of anisotropy for the natural alphabet at all chain lengths.

5.5. Classification

We start with a definition of a random fractal landscape, taken directly from Sorkin[105], which is built on the idea of fractional Brownian motion[106].

Definition. A landscape is *fractal* if $\langle \|f(x) - f(y)\|^2 \rangle \propto d^{2h}(x,y)$, where the two configurations x and y with normally distributed fitnesses $f(x)$ and $f(y)$, respectively, are separated by a distance $d(x,y)$.

As a practical matter, the distance between configurations for any finite configuration space is bounded by the diameter of the configuration space diam $\mathcal{C}$, so that Sorkin's definition makes sense only for distances small compared to this upper bound. Sorkin's definition can be reformulated in terms of autocorrelation functions:

$$1 - \rho(d) = (1 - \rho(1))d^{2h}. \tag{25}$$

For distances d small compared to the correlation length ℓ, almost all of the landscapes listed in table 1, as well as most of the other RNA landscapes, satisfy

$$\rho(d) = 1 - \frac{d}{\lambda} + \mathcal{O}\left((d/\lambda)^2 \right), \tag{26}$$

and therefore approximately satisfy Sorkin's definition with $h = 1/2$. The rescaled autocorrelation function is then of the form

$$\rho^*(\xi) = 1 - \frac{\text{diam } \mathcal{C}}{\lambda}\xi + \mathcal{O}(\xi^2) = 1 - \xi/\lambda^* + \mathcal{O}(\xi^2). \tag{27}$$

To the degree that the linear approximation holds, Sorkin's definition is satisfied. It has been suggested to use the limiting function

$$\rho^*_\infty(\xi) = \lim_{\text{diam } \mathcal{C} \to \infty} \rho^*(\xi) \tag{28}$$

as the basis for a classification of fitness landscapes. A description of classification is shown in box 11. A classification[84] depending on the form of the autocorrelation function is given in figure 10. We remark that, in the same approximation in which class **II** landscapes are fractal, they are also AR(1): $(1-q)^\xi = 1 - q\xi + \mathcal{O}(\xi^2)$. One may express this as

Lemma. An AR(1) landscape is fractal iff the correlation length is proportional to the diameter of the landscape.

I $\rho^*_\infty(\xi)$ is discontinuous for $\xi \to 0$. This corresponds to "extremely rugged" landscapes. One may divide this class into two subclasses:

Ia $\rho_\infty(\xi) = \delta(\xi, 0)$ (or more generally consists of a finite sum of delta functions). This subclass may be divided further depending on the behaviour of the unscaled correlation function:

- The unscaled correlation length ℓ diverges as diam $\mathcal{C} \to \infty$. An example of this subclass is the TSP with canonical transpositions defining the neighborhood. The correlation length is $\mathcal{O}(n)$, whereas the diameter of the landscape is $\mathcal{O}(n^2)$ where n is the number of cities; see table 4.
- The unscaled correlation length ℓ is uniformly bounded from above. Examples of this subclass are the random energy model[74], and (probably) the kinetic landscapes for the **GC**-alphabet[82].

Ib $\rho^*_\infty(\xi)$ does not vanish in any neighbourhood of $\xi = 0$ but the autocorrelation function is bounded from above by $\gamma < 1$ for $\xi \in U \setminus \{0\}$, i.e., there is a finite size step at $\xi = 0$, but the landscape is not uncorrelated. An example is provided by the asymmetric TSP with 2-opt moves, see figure 8.

II $\rho^*_\infty(\xi) = 1 - \alpha\xi + \mathcal{O}(\xi^2)$, for $\alpha = \mathcal{O}(1)$. Examples are almost all entries in table 4. Class **II** landscapes are "trivial" random fractals with $h = \frac{1}{2}$.

III $\rho^*_\infty(\xi) = 1 - \beta\xi^2 + o(\xi^2)$, for $\beta = \mathcal{O}(1)$. In contrast to all previous cases the autocorrelation function is differentiable at $\xi = 0$. These landscapes form the second class of trivial fractals, corresponding formally to Sorkin's definition of fractals with $h = 1$. These are the smoothest possible landscapes, since any function $f(x) = 1 - \mathcal{O}(x^{2+q})$ with $q > 0$ cannot be autocorrelation function of a stochastic process. Although such landscapes can be constructed no combinatorial optimization problem with a landscape of this kind in known.

IV $\rho^*_\infty(\xi) = 1 - c_p\xi^p + o(\xi^p)$ with $p \neq 1, 2$, correspond to non-trivial random fractals. Such landscapes have been constructed by Sorkin[105] but have not (yet?) been found in "real life" examples.

Figure 10: Classification of landscapes.

5.6. Ruggedness and Dynamics — Open Questions

One of the most important characteristics of a landscape is the number $\mathcal{N}$ of local optima. Palmer[107] used an at-least-exponetial increase of $\mathcal{N}$ with the system size n as definition of "rugged" landscape. Estimates for the number of local optima are known for the Nk-model[82,91,108], Sherrington-Kirkpatrick model[109], for the Travelling Salesman Problem with Transpostion metric[83], and for RNA free energy landscapes[91]. The data are consistent with the crude estimate that (for Gaussian landscapes) there is one local optimum within a ball with the radius of roughly the correlation length ℓ. While the number of local optima itself increases exponentially the probability to hit a local optimum at random, $p_{\text{l.o.}} = \mathcal{N}/N$, decreases exponentially with n on correlated (type **II**) landscapes.

A related statement is that the length of a gradient walk should be approximately the correlation length, while adaptive walks should be longer by a factor that depends on the structure of the configuration space only. Numerical evidence for the Nk model and RNA landscapes[81,91] is consistent with this statement. An example is the generalized Nk model over an alphabet with κ letters, for which holds[91]

$$\begin{aligned} \ell &= \frac{n}{k-1} \\ L_{\text{grad}} &= \frac{\kappa-1}{\kappa}\frac{\log[(\kappa-1)(k+1)]}{\log\kappa}\cdot\frac{n}{k-1} \\ L_{\text{adap}} &= \log[(\kappa-1)(k+1)]\cdot\frac{n}{k-1} \end{aligned} \tag{29}$$

The statistics of walks and local optima of an uncorrelated landscape (a random energy model) are known in detail[4,5,110,111]. This landscape is qualitatively different from the correlated landscapes discussed above in that the walk lengths are $\mathcal{O}(\log n)$ as opposed to $\mathcal{O}(n)$, and the probaility of hitting a local optimum at random is $p_{\text{l.o.}} = 1/(1+\deg)$, where deg is the degree of the a configuration, i.e., the number of nearest neighbors. A rigorous theory linking the correlation structure of Gaussian landscape with the number $\mathcal{N}$ of local optima is still missing.

Local optima are obstacles for an optimization algorithm. It is a very resonable conjecture, therefore, that optimization by a general algorithm should be easier if there are fewer local optima, or probably equivalently, if the nearest neighbor correlation is larger. Apart from the reasoning above there is some direct numerical evidence for this statement:

- The Lin-Kerninghan heuristic[113], which is based on reversals (two-opt moves) does not perform well in case of asymmetric costs. In this case the landscape is

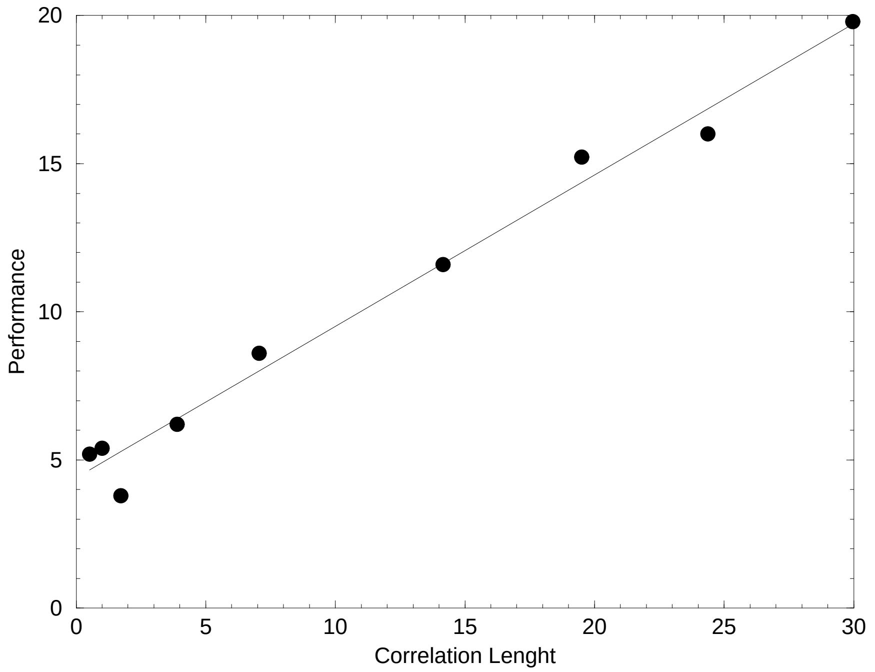

Figure 11: Performance of the genetic algorithm *Genitor* on Nk landscapes with n=96. Performance is measured by the number of improvement during 2048 generations. Data are taken from table 1 in reference [112]

class **I** in contrast to the highly correlated class II landscape of the symmetric case for which the algorithm was designed.

- Fontana finds in his thesis[114] that a quasipecies simulation using a symmetric TSP landscape yield fitter variants if reversals instead of transpositions are used as move set. Both landscapes are class **II**, but the scaled correlation length ℓ^* for reversals is twice the value for the transpositions.
- The performance of a simple genetic algortihm has been shown to depende strongly on the correlation length of the underlying landscape, see figure 11.

At present there is no satisfactory theory linking the performance of an algorithm to the structure of the landscape. And how shall we measure and compare the performance of different algorithms in the first place?

All the above results refer to Gaussian (or nearly Gaussian) landscapes. Distorting such a landscape f by a monotonic function Φ leads to a new landscape $g = \Phi(f)$ which has the same optima and the same adaptive and gradient walks as f. It has been shown, however, that the autocorrelation of the landscape

$$g(x) = \exp\left(-\beta f(x)\right) \tag{30}$$

is given by

$$\rho_g(\mu) = \frac{\alpha^{\rho_f(\mu)} - 1}{\alpha - 1} \qquad \text{with} \quad \alpha = \mathrm{e}^{\beta^2\sigma^2} \geq 1 \tag{31}$$

when f is Gaussian with variance σ and correlation function ρ_f. It is easy to check that $|\rho_g| \leq |\rho_f|$ (with equality only if $\rho_f = -1, 0, +1$), and that $\rho_g(\mu) \to \delta_{0\mu}$ for $\beta \to \infty$, i.e., the landscape can be distorted such that the correlations vanish while leaving the statistics of optima and walks unchanged.

The relation between the critical mutation rate p_{crit}, i.e., the error threshold, and the structure of a landscape is of utmost importance for evolutionary biotechnology. Evolutionary adaptation is most efficient if operating only slighly below the critical mutation rate. The possibility of predicting the optimal replication accuracy is therefore a central question for the theory of landscapes. McCaskill derived expressions for uncorrelated landscapes with a variety of fitness distributions[115]. For landscapes g of the form (30) on Boolean hypercubes a rough estimate for the errorthreshold is given by

$$p_{\mathrm{cr}} \approx \ell^* \beta d_0 f(o) n^{-2}, \tag{32}$$

where ℓ^* is the scaled correlation length of f, d_0 is constant close to 1 and $f(o)$ is the fitness of the global optimum[49]. An analogous result for Gaussian (or more general) landscapes is still unkown.

We close this section with a brief discussion of an alternative line of investigation relating the dynamics of (optimization) processes to the structure of landscapes. The set $\{x \in \mathcal{C} | f(x) \leq \epsilon\}$ is the excursion set at level ϵ; its connected components are known as *cycles*. Note that the two cycles $\mathbf{c}_1$ and $\mathbf{c}_2$ are either disjoint or one contains the other. Cycles hence impose a hierarchical structure on the landscape. The boundary of a cycle bd$\mathbf{c}$ consists of all configurations which are not contained in $\mathbf{c}$ themselves, but have a nearest neighbor which lies in $\mathbf{c}$. For each cycle one defines

$$F(\mathbf{c}) = \min_{x\in\mathbf{c}} f(x) - \min_{x\in\mathcal{C}} f(x) \qquad D(\mathbf{c}) = \max_{x\in\mathbf{c}} \min_{y\in\mathrm{bd}\mathbf{c}} [f(y) - f(x)]_+ . \tag{33}$$

$F(\mathbf{c})$ can be viewed as a generalization of the cost function to cycles instead of single configurations. $D(\mathbf{c})$ is called the *depth* of a cycle. Let $\mathcal{Q}^*$ denote the set of all cycles which do *not* contain a global optimum (we do not require that the global optimum is unique). Recent work on the convergence of simulated annealing[116] has revealed that the parameters

$$\mathcal{D} = \max_{\mathbf{c}\in\mathcal{Q}^*} D(\mathbf{c}) \quad \text{and} \quad \alpha = \min_{\mathbf{c}\in\mathcal{Q}^*} \frac{F(\mathbf{c})}{D(\mathbf{c})} \tag{34}$$

are closely related to the performance of the algorithm. For instance, an asymptotically good cooling schedule fulfills $kT_k \sim \mathcal{D}$, and the optimal speed of convergence is determined by α, which is sometimes called the difficulty of the landscape, see [117] and the references therein. The expected values of these parameters for the model landscapes in table 3, and their relevance for other optimization algorithms, including the quasispecies approach, are still unknown.

6. Molecular Genotypes and Phenotypes

6.1. Combinatory Maps

The dichotomy between genotypes and phenotypes is basic to biological evolution. Molecular biology has shown that the genetic information for the unfolding of organisms (which represent the phenotypes) is contained in the DNA (which is the genotype). RNA molecules in cell-free evolution experiments are both genotypes and phenotypes. As we see in figure 12 the genotype is again a sequence of nucleotide bases, this time in the form of an RNA molecule. Following Sol Spiegelman we consider the spatial structure of the RNA molecule as its phenotype. Indeed, this structure is evaluated by the selection process.

The mapping from genotypes, i.e., sequences, to phenotypes, i.e., secondary structure graphs, and further-on to properties of the phenotype defining its fitness are schematically depicted in figure 12.

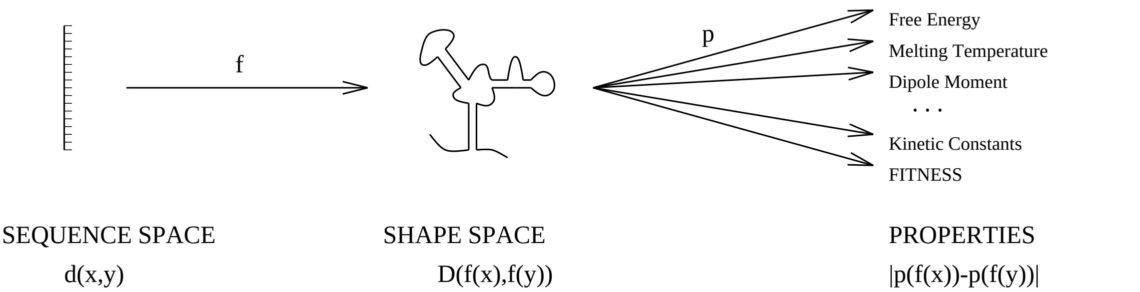

Figure 12: Schema of RNA folding.

The structure of the RNA molecule acts here as mediator between the sequence and the (scalar) property that is considered in the landscape. The relation between sequences and structures thus seems to be basic to all landscapes of biopolymers. A straightforward statistical analysis of RNA secondary structures was performed recently[24]. Structures are non scalar objects, and there is no "structure landscape". Nevertheless, RNA secondary structures form a metric space $\mathcal{Y}$ by means of one of the distance measures mentionend in section 2.4. We call a map

$$f:\ \mathcal{C} \rightarrow \mathcal{Y} \tag{35}$$

from the configuration space $\mathcal{C}$ into another metric space $\mathcal{Y}$ a *combinatory map*. Of course, a landscape is special combinatory map. The conceptual difference is, however, that landscapes have ordered sets of values, i.e., we have mountains and valleys, local optima and slopes. None of these concepts exists in a general combinatory map — we are left with nothing but equality and similarity.

6.2. RNA Shape Space

The *shape space* $\mathcal{Y}$ is defined as the set of all possible RNA secondary structures. The notion of a shape space was used previously in theoretical immunology for the set of all structures presented by all possible antigens[118,119].

The number of secondary structures which are acceptable as minimum free energy structures of RNA molecules can be computed from combinatorics of structural elements[120−123]. It is much smaller than the number of different sequences, since we have

$$S_n \sim 1.485 \times n^{-3/2}(1.849)^n \tag{36}$$

different secondary structures for κ^n sequences, i.e., there are even less structures than there are binary sequences.

The mapping from sequence space into shape space cannot be inverted: many sequences have to fold into the same secondary structure. We find that the frequency distribution of sequences folding into a common structure follows a generalized Zipf's law

$$\phi(r) = a(r+b)^{-c} \tag{37}$$

where r is the rank of a secondary structure S by the number of sequences folding into this particular structure and $\phi(r)$ is the fraction of sequences folding into S. The parameters b and c describe the number of very frequent structures and the fall-off of the tail of very rare structures, respectively. a is a normalization factor. An example is shown in figure 13. The frequency distribution of structures has a very sharp peak: relatively few structures are very common, many structures are rare and play no statistically significant role.

The distribution of secondary structures cannot be computed directly for long chains ($n \geq 50$) because of the larger number of different structures. One can, however, use a coarse grained representation of the structure. This yields again Zipf's law like rank-order statistics[121].

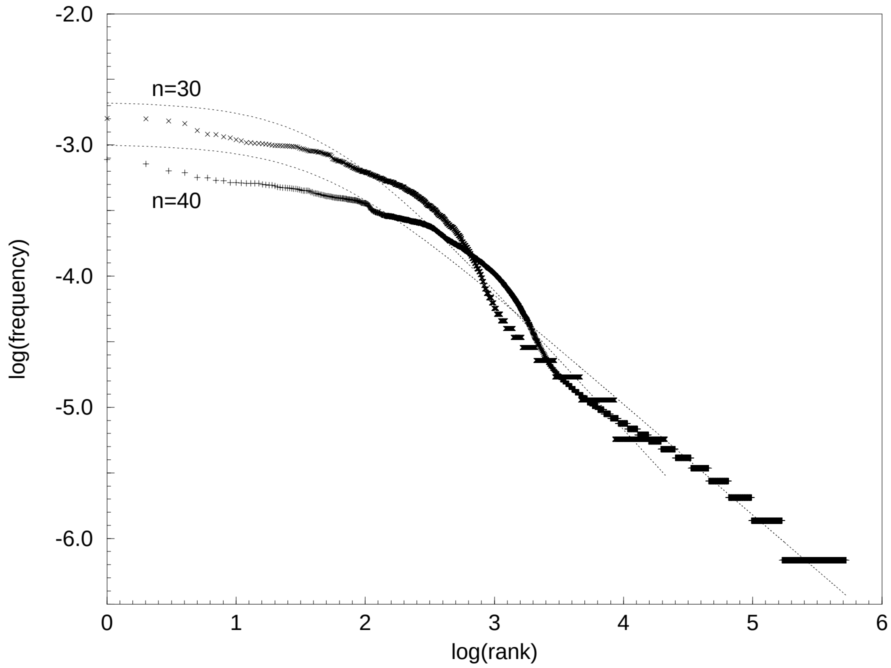

Figure 13: Zipf's law like distribution of sequences folding into common secondary structures. Data are for **GCAU** alphabet with chainlength n=30 and n=40. The open structure is not included in the data. It occurs with a frequency of a few precent at these chain lengths. The dotted lines are least square fits to equation (37). The bump at intermediate ranks is caused by **GC** rich sequences which show a much steeper tail in their structure distribution[29].

6.3. Characterization of Combinatory Maps

A theory of combinatory maps analogous to the theory of landscapes discussed in section 5 can be constructed based on equ. (17). We use

$$\begin{aligned} \rho(\mu) &= 1 - \frac{\mathcal{E}[D^2(f(x),f(y))]}{\mathcal{E}[D^2(f(p),f(q))]} \qquad \text{for } (x,y)\in\mu \text{ and } (p,q)\in\mathcal{C} \\ \hat{\rho}(\mu) &= 1 - \frac{\langle D^2(f(x),f(y))\rangle_{(x,y)\in\mu}}{\langle D^2(f(p),f(q))\rangle} \end{aligned} \tag{38}$$

as the *definition* of the autocorrelation function. The fact that there are no mean values implies that we may define a landscape as isotropic if $\mathcal{E}[D^2(f(x),f(y))]$ is constant on a symmetry class μ. The notion of empirical isotropy also carries over: the squared differences in equation (17) are simply replaced by the corresponding squared distances. In the same spirit we may define unscaled and scales structure correlation lengths ℓ and ℓ^*, respectively.

As we have seen in section 5.6, the correlation length is a statistical measure of the hardness of optimization problems. The structure correlation length allows for a similar interpretation. The shorter the correlation length, the more likely is a structural change occurring as a consequence of mutation. The correlation length thus measures stability against mutation.

Table 5. Asymptotic behaviour of correlation lenght for RNA secondary structurs.

	ΔG			
Alphabet	`MFE`	`MFE[s]`+	`PF`	`KIN`
GC	0.0222	0.0245	~ 0.02	0.0224
AU	0.0296			
GU	0.0479			
AUI	0.0299			
AGC	0.1414			
AUC	0.0792			
AUG	0.0768			
UGC	0.3135			
GCAU	0.0524	0.0414	~ 0.15	0.0616
GCAU%	0.0659			
GCXK	0.0491			
ABCDEF	0.0793			

+ See table 4.
% **GU** pairs forbidden.

Example. An almost linear increase of the strucure correlation length ℓ with chain length is observed for RNA sequences. Substantial differences are found in the correlation lengths derived from different base pairing alphabets. In particular, the structures of natural **AUGC** sequences are much more stable against mutation than pure **GC**-sequences or pure **AU**-sequences. This observation is in agreement with structural data obtained for ribosomal RNA[124]. It provides also a plausible explanation for the use of two base pairs in nature: optimization in an RNA world with only one base pair would be very hard, and the base pairing probability in sequences with three base pairs is rather low and hence most random sequences of short chain lengths ($n < 50$) do not form thermodynamically stable structures. The choice of two base pairs thus appears to be a compromise between stability against mutation and thermodynamic stability. An alternative explanation for the usage of two base pairs in nature was published recently[125,126]: the current alphabet is understood to be optimal in an RNA world where replication fidelity decreases and catalytic efficiency increases with alphabet size. Both hypothesis are experimentally testable and hence we may expect a decision in favor of one of the two alternatives in the future.

More detailed information on a combinatory map can be obtained from the conditional probabilities

$$P(D, \mu) \;=\; \text{Prob}\left\{\, D(f(x), f(y) = D \;\middle|\; (x, y) \in \mu \,\right\}. \tag{39}$$

that the images of two randomly chosen configuration have a distance D in shape space *given* that the pair of configurations (x, y) belongs to the symmetry class μ in configuration space. We call $P(D|\mu)$ a *density surface*. The autocorrelation function can be calculated from a density surface using

$$\begin{aligned}
\langle D^2(f(x), f(y))\rangle_{(x,y)\in\mu} &= \sum_D D^2\, P(D|\mu) \\
\langle D^2(f(p), f(q))\rangle_{p,q\in\mathcal{C}} &= \sum_\mu p(\mu)\langle D^2(f(x), f(y))\rangle_{(x,y)\in\mu}
\end{aligned} \tag{40}$$

and equation(38).

Example. The structure density surface (SDS) for natural **AUGC**-sequences and for pure **GC** sequences of chain length $n = 100$ is shown in figure 14. The SDS of **AUGC**-sequences has an overall shape that corresponds to one half of a horseshoe and rugged details superimposed on it. The contour plot illustrates an important property of the structure density surface: at short Hamming distances $(1 \leq n < 16)$ the probability density changes strongly with increasing Hamming distance, but further away from the reference sequence $(16 < n < 100)$ this probality density is essentially independent of the Hamming distance h. The first part reflects the local features of sequence-structure relations. Up to a Hamming distance of $h = 16$ there is still some memory of the reference sequence. Then, at larger Hamming distances, the structure density surface contains exclusively global information which is independent of the reference. The critical Hamming distance at which we observe the change from the local to the global probability distribution is a little more than twice the correlation length of the tree distance $(\approx 2.1 \times \ell_t)$.

The density surfaces contain of course more information than the autocorrelation function. For instance, the probability for finding a neutral neighbor, i.e., for finding a mutant that folds into the same structure, is given by $p_{\rm NN} = P(0|1)$. Figure 14 indicates that there is a fairly large probability for finding a neutral neighbor, $p_{\rm NN} \approx 0.4$ for the natural **GCAU**-alphabet for $n = 100$.

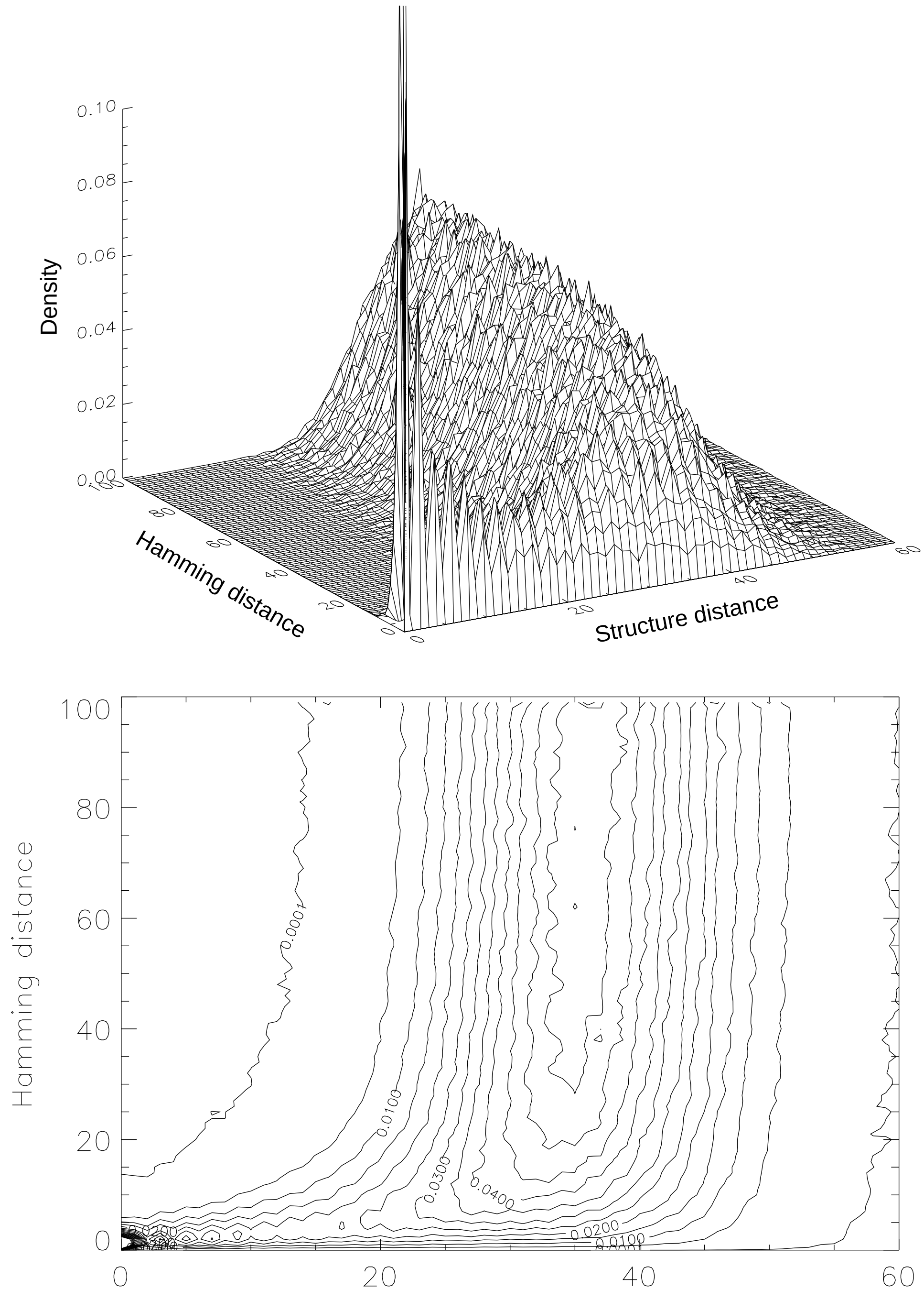

Figure 14: The structure density surface $P(D|d)$ of natural **AUGC**-sequences of chain length n=100. The density surface (upper part) is shown together with a contour plot (lower part). In order to dispense from confusing details the contour lines were smoothened. In this computation a sample of 1000 reference sequences was used which amounts to a total sample size of 10^6 individual RNA foldings.

6.4. The Inverse Folding Problem

In order to gain more information on the relation between RNA sequences and structures an inverse folding algorithm which determines the sequences that share the same minimum free energy secondary structure was conceived and applied to a variety of different structures[121].

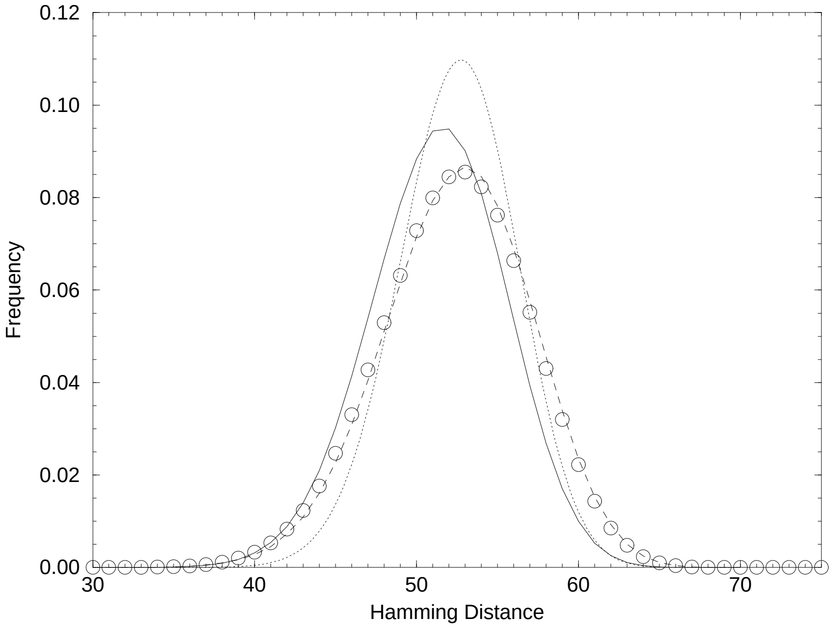

Figure 15: Distribution of 2621 tRNA analoga obtained from reverse folding ○. The start sequences did not contain **GU** pairs, and only very few **GU** pairs are found in the rRNA analoga. For comparison the distance distribution functions for random sequences (dotted line), random compatible sequences (dashed line), and random compatible sequences without **GU** pairs are shown. This indicates that sequences are randomly distributed (subject to the constraint of being compatible with the stucture). This result was confirmed using statistical geometry[41] and split decomposition[127].

In addition these results show that sequences folding into the same secondary structure are, in essence, randomly distributed. Figure 15 presents an example of sequences folding into he secondary structure of phenylalanyl-tRNA. Practically identical results were obtained for other, arbitrarily chosen secondary, structures. Since there are relatively few common structures and the sequences folding into the same structure are randomly distributed in sequence space, all common structrues are found in relatively small patches of sequence space. For natural **AUGC**-sequences of chain length $n=100$ a sphere of radius $h=16$ (in Hamming distance) is sufficient to yield already the global distribution of structure distances. In this case we can expect all common structures to be found in such spheres in sequence space. There are as many as 6.2×10^{25} sequences in such a ball. Although this number is large, it is nothing compared to the total number of sequences of this chain length: $4^{100} \approx 1.6 \times 10^{60}$.

6.5. Neutral Networks

We have seen that the sequences $U(S)$ folding into the same secondary structure S are randomly distributed in sequence space. On the other hand there is a high probability for finding neutral neighbors. Hence the sequences $U(S)$ are not isolated in sequence space. In order to refine our understanding of the geometrical arrangement of $U(S)$ in sequence space a computer experiment was carried out allowing for an estimate of the diameter of connected sets of neutral sequences. We search for "neutral paths" through sequence space. The Hamming distance from the reference increases monotonously along such a neutral path but the structure remains unchanged. A neutral path ends when no further neutral sequence is found in the neighbourhood of the last sequence. The length $\mathcal{L}$ of a path is the Hamming distance between the reference sequence and the last sequence, and hence a lower bound on the diameter of the connected "neutral network". Clearly, a neutral path cannot be longer than the chain length, $\mathcal{L} \leq n$. The length distribution of neutral paths in the sequence space of RNA molecules of chain length $n = 100$ is shown in figure 16 for the the natural **GCAU**-alphabet.

It is remarkable that about 20 % of the neutral paths in sequence space of natural **AUGC** sequences have the maximum length. They lead through the entire sequence space to a sequence differing in all positions from the reference but still sharing its structure. In shape spaces derived from binary sequences almost no neutral path reaches the complementary sequence. This is certainly a consequence of the symmetry of the binomial distribution: there are very few sequences in the error classes $n-1$, $n-2$, etc., and it is unlikely that we find one among them which folds precisely into the same structure as the reference sequence. Still the average lenght $\mathcal{L}$ is much larger than the average distance of two randomly chosen sequences, $\mathcal{L} \gg n/2$. For details see ref.[121].

The union of all neutral paths forms a dense neutral network in the example considered here. This, of course, need not be the case in general: we may have short neutral paths confined to small disjoint regions in sequence space. We may expect

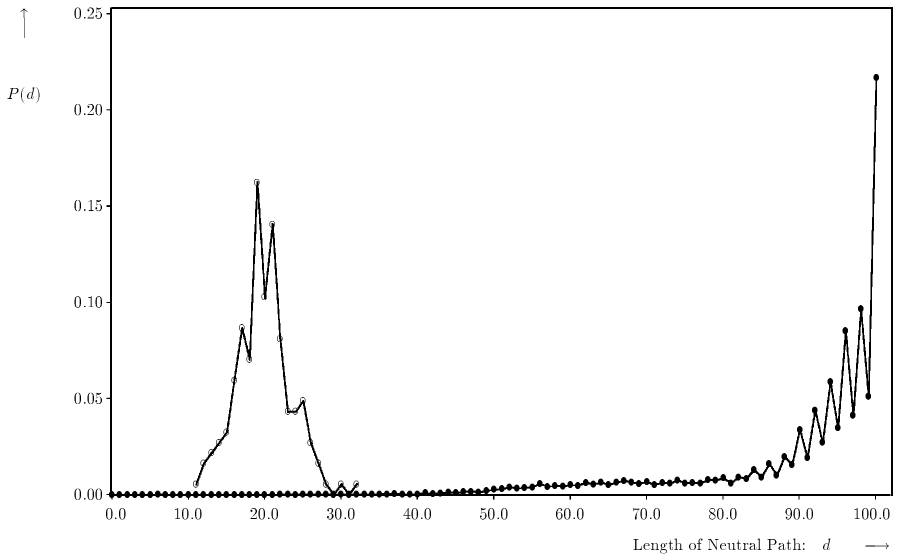

Figure 16: Neutral Path and Closest Approach. Two probability densities a function of Hamming distance are shown in this figure:
(i) the shortest distances between trial and target sequences (○) obtained as endpoints of neutral paths *approaching* the target from a random trial sequence (185 targets and 100 trials for each target), and (ii) longest distances between the reference and the end points of monotonously diverging neutral paths ● (500 reference sequences were used).

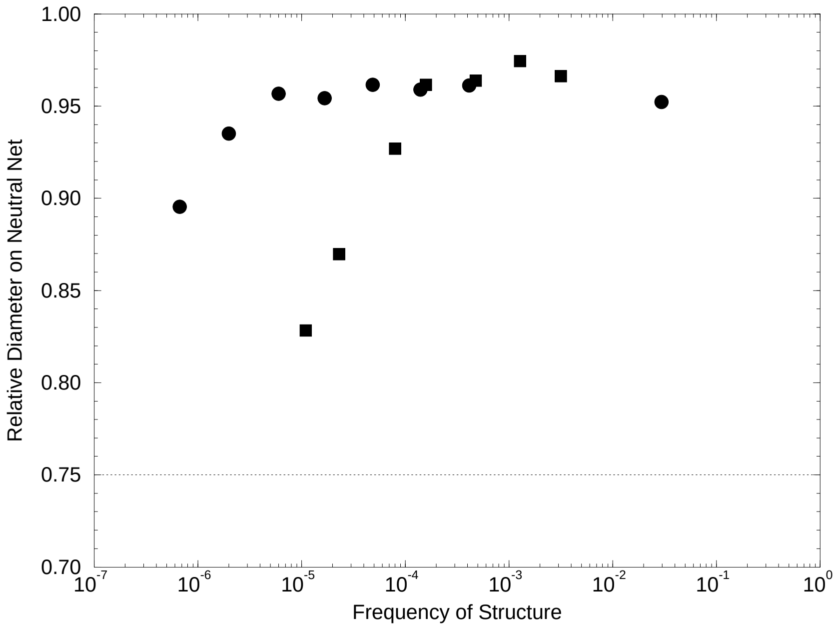

Figure 17: Average length of neutral path as a function of the frequency of the underlying structure. □ n=30, • n=40. The data shown here are lower bounds on the average diameter of the neutral networks.

a chracteristic change in the hardness of the optimization problem depending on whether the network of neutral paths is below or above a percolation threshold. It is worth noting that neutral nets are not a peculiarity of the few most frequent structures. As shown in figure 17 even the rarest structures we were able to find give rise to networks that reach way beyond the average distance of random sequences. The data indicate, however, a sharp (chain length dependent) transition at which the average diameter of neutral networks begins to decrease with decreasing frequency of the underlying secondary structure. The percolation properties of the sequence-structure map deserve further investigation.

6.6. Summary

The evidence presented above allows us to make the following four statements:

(1) Sequences folding into one and the same structure are distributed randomly in sequence space,

(2) The frequency distribution of structures is sharply peaked (there are comparatively few common structures and many rare ones),

(3) Sequences folding into all common structures are found within (relatively) small neighbourhoods of any random sequence, and

(4) The shape space contains extended "neutral networks" joining sequences with identical structures (a large fraction of neutral path leads from the initial sequence through the entire sequence space to a final sequence on opposite side – there are $(\kappa-1)^n$ sequences which differ in all positions from an initial sequence).

Combining the two statements (1) and (3) we may visualize the mapping from sequences into structures as illustrated by the sketch shown in figure 18.

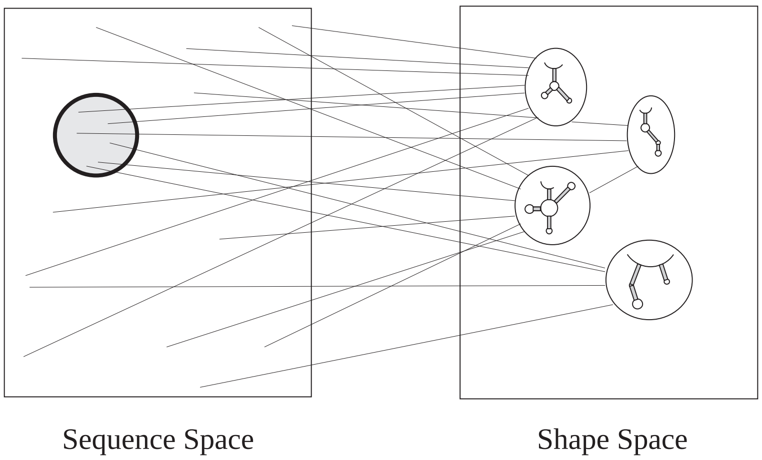

Figure 18: A sketch of the mapping from sequences into RNA secondary structures as derived here. Any random sequence is surrounded by a ball in sequence space which contains sequences folding into (almost) all common structures. The radius of this ball is much smaller than the dimension of sequence space.

7. Implications

7.1. Biotechnology

The capability of RNA molecules to evolve by natural selection is used nowadays in applied molecular evolution to produce biomolecules with new properties. This novel evolutionary technology was originally suggested by Eigen and Gardiner[128]. Somewhat later Kauffman suggested the use of random sequences in selection experiments[129]. One approach to the problem is suitable for "batch" experiments (Figure 17). The essential trick of this technique is to encode the desired functions into the selection constraint. As indicated in figure 17 sufficient variation is introduced into populations either by artificially increased mutation rates or by partial randomization of sequences. The synthesis of oligonucleotides with random sequences is routine nowadays. Several examples of successful applications of molecular selection techniques to biochemical problems are found in the current literature[130−136].

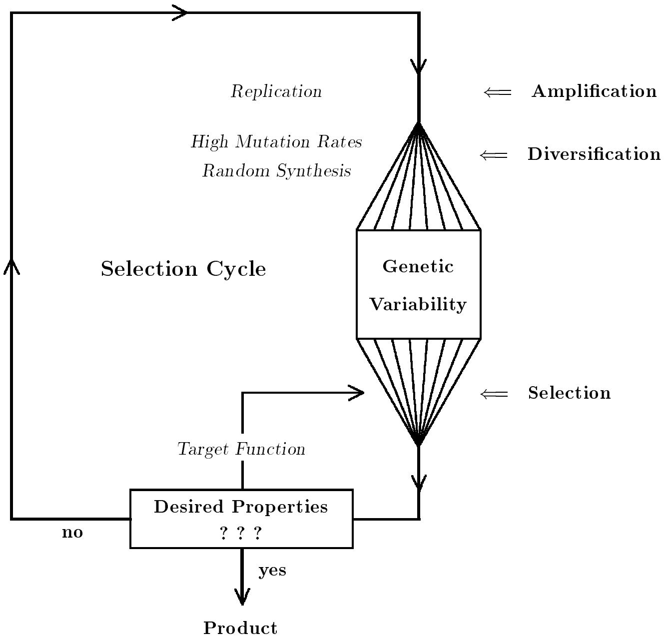

Figure 19: A selection technique of applied molecular evolution based on encoding of the function to be developed into the selection constraint.

In reality it will often be impossible to encode the desired function directly into the selection constraint. Then spatial separation of individuals and massively parallel screening provides a solution (Figure 19). This technique, however, requires highly sophisticated equipment which is currently under development[57,60,137].

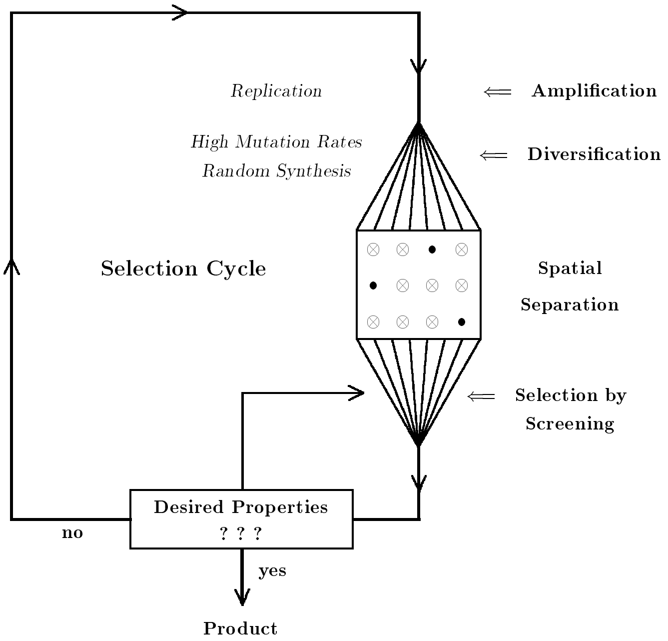

Figure 20: A selection technique of applied molecular evolution based on massively parallel screening.

The results on RNA shape space derived in the previous section suggest straightforward strategies in the search for new RNA structures. It is of little advantage to start from natural or other preselected sequences since any random sequence would serve equally well as the starting molecule for the selection cycles shown in figure 20. Any common secondary structure with suitable properties is accessible within a few selection cycles. Since the secondary structure is only a crude first approximation to the actual 3D-structure, fine tuning of properties by choosing from a variety of molecules sharing the same secondary structure will be necessary. In order to achieve this goal it is advantageous to adopt alternations of selection cycles with low and high error rates. At low error rates the population performs a search in the vicinity of the current master sequence (the most common sequence which usally is the fittest sequence as well). If no RNA molecule with satisfactory properties is found, a change to high error rate is adequate. Then the population spreads along the neutral network to other regions of sequence space which are explored in detail after tuning the error rate low again.

7.2. An RNA World?

The structure of shape space is highly relevant for evolutionary optimization in nature too. Since long neutral paths are common, populations drift readily through sequence space whenever selection constraints are absent. This is precisely what is predicted by the "neutral theory of evolution"[45], and what is observed in molecular phylogeny by sequence comparisons of different organisms. The structure of shape space provides also a firm answer to the old probability argument against the possibility of successful adaptive evolution[138]. How should nature find a given biopolymer by trial and error when the chance to guess it is as low as $1{:}\kappa^n$? Previously given answers[30,31] can be supported and extended by precise data on the RNA shape space. The numbers of sequences that have to be searched in order to find adequate solutions are many orders of magnitude smaller than those guessed on naive statistical grounds. If one of the common structures has a property which increases fitness it can hardly be missed in an evolutionary search.

RNA molecules were found to be unique in chemistry and biology since they can not only act as templates for their reproduction but are also catalysts for several reactions involving other RNA molecules or oligopeptides. The discoveries of Thomas Cech[139] and Sidney Altman[140–142] have shown that, in contrast to previous belief, RNA molecules can catalyze several classes of reactions with similar high efficiency and specificity as protein enzymes. The notion of "ribozymes" was created for such RNA based catalysts. The catalytic activities of RNA molecules are all related to cleavage or formation of nucleotide or peptide bonds. For RNA substrates ligation and cleavage are highly sequence specific. The capability of RNA to catalyze peptide bond cleavage and formation[143–145] is of particular interest since it can be interpreted as a hint that primordial ribosomes might have worked without proteins. Although the catalytic repertoire of ribozymes is rather poor compared to that of protein enzymes, it comprises the key reactions that are necessary to produce polynucleotides from oligomers and monomers. It is generally believed nowadays that there was a period in the origin of life on Earth when RNA served as both, genetic material and specific catalyst[146,147].

Is there room for an RNA world in the current view of chemical evolution on the primordial Earth? A plausible sequence of scenarios leading to the onset of RNA based Darwinian evolution is discussed in refs.[148,149]. The major problem with RNA in chemical evolution is to be seen in its highly elaborate molecular structure: for RNA template induced reactions molecules of high stereochemical purity are required, and chiral compounds such as ribose have to be present as pure single antipods (Racemic mixtures of D- and L-ribose incorporated into polynucleotides are prohibitive for template induced replication). It was suggested therefore that less complicated molecules which could nevertheless act as templates preceded the RNA world[150,151]. Recent progress in template chemistry[6] makes this suggesetion more and more plausible. RNA molecules, on the other hand, are the first template molecules which share most of their chemical properties with DNA the predominant genetic information carrier at present. Information transfer from DNA to RNA and vice versa, known as transcription and reverse transcription, is routine in present day biochemistry. We can speculate therefore that a phylogenetic extrapolation of present day genetic information may go back as far as to the first efficiently replicating species of a primordial RNA world. In short RNA molecules seem to appear late in chemical evolution and very early in biological evolution.

8. Acknowledgements

The statistical analysis of the RNA shape space presented here is joint work with Walter Fontana and Ivo L. Hofacker. It will be described in detail elsewhere. Financial support by the Austrian *Fonds zur Förderung der wissenschaftlichen Forschung* (Projects S 5305-PHY, P 9942-PHY and P 8526-MOB) and by the Commission of the European Communities (Contract PSS*0396) is gratefully acknowledged. The *Institut für Molekulare Biotechnologie (Jena, Germany)* is sponsored by core funding from the *Thüringische Ministerium für Wissenschaft und Kunst (Erfurt)* and the *Bunderministerium für Forschung und Technologie (Bonn)*. PFS thanks the German *Max Planck Gesellschaft* for supporting his stay at the Santa Fe Institute. We thank Dr. Walter Grüner for reading the manuscript.

9. References

[1] Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth International Congress on Genetics* **1** 356-366.

[2] Kauffman, S.A. (1989). Adaptation on Rugged Fitness Landscapes. *Complex Systems.* Santa Fe Institute Studies in the Sciences of Complexity, D. Stein. Redwood City: Addison Wesley. 527-618.

[3] Kauffman, S.A., and S.Levin. (1987). Towards a general theory of adaptive walks on rugged landscapes. *J.Theor.Biol.* **128** 11-45.

[4] Macken, C.A., and A.S.Perelson. (1989). Protein evolution on rugged landscapes. *Proc.Natl.Acad.Sci.USA* **86** 6191-6195.

[5] Macken, C.A., P.S.Hagan, and A.S.Perelson. (1991). Evolutionary walks on rugged landscapes. *SIAM J.Appl.Math.* **51** 799-827.

[6] Orgel, L.E. (1992). Molecular replication. *Nature* **358** 203-209.

[7] Riesner, D., and Gross, H.J. (1985). Viroids. *Ann. Rev. Biochem.* **54** 531-564.

[8] Biebricher, C.K., M.Eigen, and W.C.Gardiner, Jr. (1983). Kinetics of RNA replication. *Biochemistry* **22** 2544-2559.

[9] Biebricher, C.K., M.Eigen, and W.C.Gardiner, Jr. (1984). Kinetics of RNA replication: Plus-minus asymmetry and double-strand formation. *Biochemistry* **23** 3186-3194.

[10] Biebricher, C.K., M.Eigen, and W.C.Gardiner, Jr. (1985). Kinetics of RNA replication: Competition and selection among self-replicating RNA species. *Biochemistry* **24** 6550-6560.

[11] Zuker, M. (1989). The Use of Dynamical Programming in RNA Secondary Structure Prediction. In: Waterman M.S. (ed). *Mathematical Methods for DNA Sequences.* CRC Press, Boca Raton pp. 159-184.

[12] Zuker, M. and D. Sankoff. (1984). The use of dynamics programming algorithms in RNA secondary structure predictions. *Bull. Math. Biol.* **46** 591-621.

[13] Martinez, H.M. (1984). An RNA Folding Rule. *Nucl. Acids Res.* **12** 323-334.

[14] Tacker, M. (1993). Robust Properties of RNA Folding. Doctoral Thesis, University of Vienna .

[15] McCaskill, J.S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29** 1105-1119.

[16] Nussinov, R., G. Pieczenik, J.R. Griggs, and D.J. Kleitman. (1978). Algorithms for loop matching. *SIAM J. Appl. Math.* **35** 68-82.

[17] Mironov, A.A., L.P. Dyakonova, and A.E. Kister. (1985). A kinetic approach to the prediction of RNA secondary structures. *J. Biomol. Struct. Dyn.* **2** 953-962.

[18] Mironov, A.A. and A.E. Kister. (1985). RNA secondary structure formation during transcription. *J. Biomol. Struct. Dyn.* **4** 1-9.

[19] Zuker, M. (Public Domain Software). `mfold-2.2`. *Site:* `nrcbsa.bio.nrc.ca`*; directory:* `pub/mfold.tar.Z` .

[20] Hofacker I.L., Fontana W., Stadler P.F., Bonhoeffer L.S., Tacker M., and Schuster P. (Public Domain Software). `ViennaRNA-1.03`. *Site:* `ftp@itc.univie.ac.at`*; directory:* `pub/RNA/ViennaRNA-1.03.tar.Z` .

[21] Hofacker I.L., Fontana W., Stadler P.F., Bonhoeffer L.S., Tacker M., and Schuster P. (1993). Fast Folding and Comparison of RNA Secondary Structures. *Mh. Chem.* in press.

[22] Shapiro B.A. (1988). An algorithm for comparing multiple RNA secondary structures. *CABIOS* **4** 387-397.

[23] Shapiro B.A., and Zhang, K. (1990). Computing multiple RNA secondary structures using tree comparisons. *CABIOS* **6** 309-318.

[24] Fontana W., Konings D.A.M., Stadler P.F., and Schuster P. (1993). Statistics of RNA Secondary Structures. *Biochemistry* **33** 1389-1404.

[25] Tai, K. (26). The Tree-to-Tree Correction Problem. *J. ACM* **1979** 422-433.

[26] Ohmori K., and Tanaka E. (1988). A Unified View on Tree Metrics. In: *Syntactic and Structural Pattern Recognition* G. Ferrate (ed.) Springer-Verlag: Berlin, Heidelberg pp. 85-100.

[27] Hogeweg, P. & Hesper, B. (1984). Energy Directed Folding of RNA Sequences. *Nucl. Acid Res.* **12** 67-74.

[28] Konings, D.A.M. & Hogeweg, P. (1989). Pattern Analysis of RNA secondary structure, similarity and consensus of minimal-energy folding. *J. Mol. Biol.* **207** 597-614.

[29] Tacker, M., P.F.Stadler, E.G.Bornberg-Bauer, I.L.Hofacker and P.Schuster. (1993). Robust Properties of RNA Secondary Structure Folding Algorithms. *in preparation* .

[30] Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58** 465-523.

[31] Eigen, M., and P.Schuster. (1977). The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. *Naturwissenschaften* **64** 541-565.

[32] Eigen, M., J.McCaskill, and P.Schuster. (1988). Molecular quasispecies – An abridged account. *J. Phys. Chem.* **92** 6881-6891.

[33] Eigen, M., J.McCaskill, and P.Schuster. (1989). The molecular quasispecies. *Adv. Chem. Phys.* **75** 149-263.

[34] Jones, B.L., R.H. Enns, and S.S. Ragnekar. (1976). On the Theory of Selection in Coupled Macromolecular Systems. *Bull. Math. Biol.* **38** 12-28.

[35] Thompson, C.J. and J.L. McBride. (1974). On Eigen's Theory of Selforganization of Molecules and the Evolution of Biological Macromolecules. *Math. Biosc.* **21** 127-142.

[36] Hamming, R.W. (1950). Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29** 147-160.

[37] Hamming, R.W. (1980 and 1986). *Coding and information theory*. Englewood Cliffs (N.J.): Prentice Hall .

[38] Maynard-Smith, J. (1970). Natural selection and the concept of a protein space. *Nature* **225** 563-564.

[39] Rechenberg, I. (1973). *Evolutionsstrategie*. Stuttgart-Bad Canstatt (Germany): Frommann-Holzboog .

[40] Swetina, J., and P.Schuster. (1982). Self-replication with errors - A model for polynucleotide replication. *Biophys. Chem.* **16** 329-345.

[41] Eigen, M., R.Winkler-Oswatitsch, and A.Dress. (1988). Statistical geometry in sequence space: a method of quantitative comparative sequence analysis. *Proc. Natl. Acad. Sci. USA* **85** 5913-5917.

[42] Schuster, P. and K. Sigmund. (1985). Dynamics of evolutionary optimization. *Ber. Bunsenges. Phys. Chem.* **89** 668-682.

[43] Schuster, P. (1992). Complex Optimization in an Artificial RNA World. In: C.G.Langton, C.Taylor, J.D.Farmer, and S.Rasmussen (eds). *Artificial Life II*. Santa Fe Institute Studies in the Sciences of Complexity, Vol X. Addison-Wesley, Redwood City (Cal.) pp. 277-291.

[44] Schuster P., J. Swetina. (1988). Stationary Mutant Distribution and Evolutionary Optimization. *Bull. Math. Biol.* **50** 635-660.

[45] Kimura M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press: Cambridge (UK) .

[46] Nowak M.A. and P. Schuster. (1989). Error Thresholds for Replication in Finite Populations. Mutation Frequencies and the Onset of Muller's Ratchet. *J. Theor. Biol.* **137** 375-395.

[47] Derrida B. and L. Peliti. (1991). Evolution in a Flat Fitness Landscape. *Bull. Math. Biol.* **53** 355-382.

[48] Schnabl, W. (1990). Adaptive Optimization of Polynucleotides with Variable Chain Length. PhD Thesis. University of Vienna .

[49] Bonhoeffer, L.S. and P.F. Stadler. (1993). Error Thresholds on Correlated Fitness Landscapes. *J. Theor. Biol.* in press..

[50] Leuthäusser, I. (1986). An exact correspondence between Eigen's evolution model and a two-dimensional Ising model. *J. Chem. Phys.* **84** 1884-1885.

[51] Leuthäusser, I. (1987). Statistical mechanics of Eigen's evolution model. *J. Stat. Phys.* **48** 343-360.

[52] Demetrius, L., P. Schuster and K. Sigmund. (1985). Polynucleotide evolution and branching processes. *Bull. Math. Biol.* **47** 239-262.

[53] Demetrius, L. (1988). Random Spin Models of Chemical Kinetics. *J. Chem. Phys.* **87** 6939-6964.

[54] Tarazona, P. (1992). Error Thresholds for Molecular Quasispecies as Phase Transitions: From Simple Landscapes to Spinglass Models. *Phys. Rev. A* **45** 6038-6050.

[55] Spiegelman, S. (1971). An approach to the experimental analysis of precellular evolution. *Quart. Rev. Biophys.* **4** 213-253.

[56] Husimi, Y., K.Nishgaki, Y.Kinoshita, and T.Tanaka. (1982). Cellstat – A continuous culture system of a bacteriophage for the study of the mutation rate and the selection process at the DNA level. *Rev. Sci. Instrum.* **53** 517-522.

[57] Husimi, Y., and H.-C.Keweloh. (1987). Continuous culture of bacteriophage Qβ using a cellstat with a bubble wall-growth scraper. *Rev. Sci. Instrum.* **58** 1109-1111.

[58] Bauer, G.J., J.S. McCaskill, and H. Otten. (1989). Traveling waves of *in vitro*evolving RNA. *Proc. Natl. Acad. Sci. USA* **86** 7937-7941.

[59] Kramer, F.R., D.R. Mills, P.E. Cole, T. Nishihara, and S. Spiegelman. (1974). Evolution *in vitro*: sequence and phenotype of a mutant RNA resistant to ethidium bromide. *J. Mol. Biol.* **89** 719-736.

[60] Strunk, G. (1993). Automatized evolution experiments *in vitro*and natural selection under controlled conditions by means of the serial transfer technique, PhD Thesis, Universität Braunschweig (Germany). .

[61] Biebricher, C.K., and R. Luce. (1992). Sequence analysis of RNA species synthesized by Qβ replicase without template. *Biochemistry* **32** 4848-4854.

[62] Maynard-Smith, J. (1989). *Evolutionary genetics.* Oxford University Press: New York .

[63] Van Valen, L. (1973). A new evolutionary law. *Evol. Theory* **1** 1-30.

[64] Holland J.H. (1993). *Adaptation in Natural and Artificial Systems.* MIT Press, Cambridge, MA .

[65] Goldberg D.E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, Reading, MA .

[66] Lawler, E., J.Lenstra, A.Rinnooy-Kan and D.Shmoys. (1985). *The travelling salesman problem. A guided tour of combinatorial optimization.* J. Wiley & Sons, New York .

[67] Garey, M., and D.Johnson. (1979). *Computers and intractability: A guide to the theory of NP-completeness.* Freeman, San Francisco .

[68] Bland, R.G., and D.F.Shallcross. (1988). Large traveling salesmen problems arising from experiments in X-ray crystallography. *Oper. Res. Lett.* **8** 125-128.

[69] Chan, D., and D.Mercier. (1989). IC insertion: An application of the TSP. *Int. J. Prod. Res.* **27** 1837-1842.

[70] Bohr, H., and S.Brunak. (1990). Travelling salesman approach to protein conformation. *Complex Systems* **3** 9-28.

[71] Miller D.L. and J.F. Pekny. (1991). Exact Solution of Large Asymmetric Traveling Salesman Problems. *Science* **251** 754-761.

[72] Chowdhury, D. (1986). *Spin glasses and other frustrated systems.* Princeton Univ. Press, Princeton, NJ .

[73] Mezard M., G. Parisi, and M.A. Virasoro (eds). (1987). *Spin glass theory and beyond.* World Scientific, Singapore .

[74] Derrida, B. (1981). Random Energy Model: An exactly solvable model of disordered systems. *Phys. Rev. B* **24** 2613-2626.

[75] Kauffman, S.A., and S.Levin. (1987). Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.* **128** 11-45.

[76] Kauffman, S.A., and E.D.Weinberger. (1989). The Nk model of rugged landscapes and its application to the maturation of the immune response. *J. Theor. Biol.* **141** 211-245.

[77] Lau, K.F., and K.A.Dill. (1990). Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. USA* **87** 638-642.

[78] Yue, K., and K.A.Dill. (1993). Sequence structure relationships in proteins and co-polymers. *Phys. Rev. E* **48** 2267-2278.

[79] Camachp, C., and D.Thirumalai. (1993). Minimum energy compact structures of random sequences of heteropolymers. *Phys. Rev. Lett.* **71** 2505-2508.

[80] Binder, K., and A.P.Young. (1986). Spin glasses: Experimental facts, theoretical concepts, and open questions. *Rev. Mod. Phys.* **58** 801-976.

[81] Weinberger, E.D. (1991). Fourier series and Taylor series on landscapes. *Biol. Cybern.* **65** 321-330.

[82] Fontana, W., T. Griesmacher, W. Schnabl, P.F. Stadler, and P. Schuster. (1991). Statistics of landscapes based on free energies, replication and degradation rate constants of RNA secondary structures. *Mh.Chem.* **122** 795-819.

[83] Stadler P.F. and W. Schnabl. (1992). The Landscape of the Travelling Salesman Problem. *Phys. Lett. A* **161** 337-344.

[84] Weinberger E.D. and P.F. Stadler. (1993). Why *some* fitness landscapes are fractal. *J. Theor. Biol.* **163** 255-275.

[85] Mézard M. and G. Parisi. (1985). Replicas and Optimization. *J. Physique Lett.* **46** L 771-778.

[86] Stadler P.F. (1992). Correlation in Landscapes of Combinatorial Optimization Problems. *Europhys. Lett* **20** 479-482.

[87] Fu Y. and P.W. Anderson. (1986). Application of statistical mechanics to NP-complete problems in combinatorial optimization. *J. Phys. A: Math. Gen.* **19** 1605-1620.

[88] Stadler P.F. and R. Happel. (1992). Correlation Structure of the Landscape of the Graph-Bipartitioning-Problem. *J. Phys. A: Math. Gen.* **25** 3103-3110.

[89] Bernasconi J. (1987). Low autocorrelation binary sequences: statistical mechanics and configuration space analysis. *J. Physique* **48** 559-567.

[90] Weinberger E.D. (1991). Local Properties of Kauffman's Nk-Model, a Tunably Rugged Energy Landscape. *Phys. Rev. A* **44** 6399-6413.

[91] Fontana W., P.F. Stadler, E.G. Bornberg-Bauer, T. Griesmacher, I.L. Hofacker, M. Tacker, P. Tarazona, E.D. Weinberger, and P. Schuster. (1993). RNA folding and combinatory landscapes. *Phys. Rev. E* **47** 2083.

[92] Amitrano C., L. Peliti, and M. Saber. (1989). Population Dynamics in a Spin Glass Model of Chemical Evolution. *J. Mol. Evol.* **29** 513-525.

[93] Stadler P.F. (1993). Random Walks and Orthogonal Functions Associated with Highly Symmetric Graphs. *SFI preprint* 93-06-031.

[94] Sherrington D. and S. Kirkpatrick. (2975). Solvable Model of a Spin Glass. *Phys. Rev. Lett.* **35** 1792-1796.

[95] Bonhoeffer, L.S., J.S.McCaskill, P.F.Stadler and P.Schuster. (1993). RNA Multi-Structure Landscapes. A Study Based on Temperature Dependent Partition Functions. *Eur. Biophys. J.* **22** 13-24.

[96] Tacker M., W. Fontana, P.F. Stadler, and P. Schuster. (1993). Statistics of RNA Melting Kinetics. *SFI-preprint* 93-06-043.

[97] Salser W. (1977). Globin Messenger Sequences. Analysis of Base-Pairing and Evolutionary Implications. *Cold Spring Harbor Symp. Quant. Biol.* **42** 985-1002.

[98] Freier, S.M., R.Kierzek, J.A.Jaeger, N.Sugimoto, M.H.Caruthers, T.Neilson and D.H.Turner. (1986). Improved free-energy parameters for predictions of RNA duplex stability. *Biochemistry* **83** 9373-9377.

[99] Fontana, W., and P. Schuster. (1987). A computer model of evolutionary optimization. *Biophys. Chem.* **26** 123-147.

[100] Fontana, W., W. Schnabl and P. Schuster. (1989). Physical aspects of evolutionary optimization and adaptation. *Phys. Rev. A* **40** 3301-3321.

[101] Weinberger E.D. (1990). Correlated and Uncorrelated Landscapes and How to Tell the Difference. *Biol. Cybern.* **63** 325-336.

[102] Stadler P.F. (1993). Linear Operators on Correlated Landscapes. *Santa Fe Institute preprint* 93-06-042.

[103] Stadler P.F. and W. Grüner. (1993). Anisotropy in Fitness Landscapes. *J. Theor. Biol.* in press.

[104] Piccirilli, J.A., T.Krauch, S.E.Moroney and S.A.Benner. (1990). Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature* **343** 33-37.

[105] Sorkin, G. (1988). Combinatorial optimization, simulated annealing, and fractals. *IBM Research Report RC13674 (No. 61253)* .

[106] Voss, R. (1986). Characterization and measurement of random fractals. *Physical Scripta* **T13** 257-260..

[107] Palmer, R. (1991). Optimization on rugged landscapes. In: *Molecular Evolution on Rugged Landscapes: Proteins, RNA, and the Immune System.* A.S. Perelson and S.A. Kauffman (eds.). Addison Wesley, Redwood City (CA). pp. 3-25.

[108] Kauffman S.A. (1993). *The orgins of order. Oxford Univ. Press, Oxford* .

[109] Rokhsar D.S., P.W. Anderson, and D.L. Stein. (1986). Selforganization in prebiological systems: Simuations of a Model of the Origin of Genetic Information. *J. Mol. Evol.* **23** 119-126.

[110] Bak P., H. Flyvbjerg, and B. Lautrup. (1992). Coevolution in a rugged fitness landscape. *Phys. Rev. A[15]* **46** 6724-6730.

[111] Flyvbjerg H. and B. Lautrup. (1992). Evolution in a rugged fitness landscape. *Phys. Rev. A[15]* **46** 6714-6723.

[112] Manderick, B., deWeger M., and P. Spiessen. (1991). The Genetic Algorithm and the Structure of the Fitness Landscape. *Proc. of the 4th Internatl. Conf.*

on Genetic Algorithms R.K. Belew and L.B. Booker (eds). Morgan Kaufmann Inc., San Mateo CA pp.143-150.

[113] Lin, S., and B.W. Kernighan. (1973). An Effective Heuristic Algorithm for the Traveling Salesman Problem. *Oper. Res.* **21** 498-516.

[114] Fontana W. (1988). Ein Computermodell der Evolutionären Optimierung. *Doctoral Thesis* University of Vienna .

[115] McCaskill, J.S. (1984). A Localization Threshold for Molecular Quasispecies from Continuously Distributed Replication Rates. *J. Chem. Phys.* **80** 5194-5202.

[116] Kirkpatrick S., C.D. Gelatt, and M.P. Vecchi. (1983). Optimization by Simulated Annealing. *Science* **220** 671-680.

[117] Azencott R. (1992). *Simulated Annealing.* John Wiley & Sons, New York .

[118] Perelson, G.A., and G.F.Oster. (1979). Theoretical studies on clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *J. Theor. Biol.* **81** 645-670.

[119] Segel, L.A., and A.S.Perelson. (1988). Computations in shape space: a new approach to immune network theory. Theoretical Immunology, Part Two. Santa Fe Institute Studies in the Sciences of Complexity, Vol.III. A.S.Perelson. Redwood City (Cal.): Addison-Wesley pp. 321-343.

[120] Hofacker, I.L., P.Schuster, and P.F.Stadler. (1993). Combinatorics of RNA secondary structures. *SIAM J.Disc.Math.* submitted.

[121] Schuster, P., W.Fontana, P.F.Stadler, and I.L.Hofacker. (1993). From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures. *SFI Preprint 93-07-045* .

[122] Stein, P.R., and M.S.Waterman. (1978). On some sequences generalizing the Catalan and Motzkin numbers. *J.Discr.Math.* **26** 261-272.

[123] Waterman, M.S. (1978). Secondary structures of single-stranded nucleic acids. *Adv. Math. Suppl. Studies* **1** 167-212.

[124] Wakeman J.A. and B.E.H. Maden. (1989). 28 S Ribosomal RNA in vertebrates. Location of large-scale features revealed by electron microscopy in relation to other features of the sequences. *Biochem. J.* **258** 49-56.

[125] Szathmáry, E. (1991). Four letters in the genetic alphabet: a frozen evolutionary optimum? *Proc. Roy. Soc. London B* **245** 91-99.

[126] Szathmáry, E. (1992). What is the optimum size for the genetic alphabet? *Proc. Natl. Acad. Sci. USA* **89** 2614-2618.

[127] Bandelt H.J. and A.W.M. Dress. (1992). A Canonical Decomposition Theory for Metrics on a finite Set. *Adv. Math.* **92** 47-105.

[128] Eigen M. and W. Gardiner. (1984). Evolutionary molecular engineering based on RNA replication. *Pure & Appl. Chem.* **56** 967-978.

[129] Kauffman S.A. (1986). Autocatalytic sets of proteins. *J.theor.Biol.* **119** 1-7.

[130] Joyce, G.F. (1992). Directed molecular evolution. *Sci. Am.* **267(6)** 48-55.

[131] Kauffman, S.A. (1992). Applied molecular evolution. *J. Theor. Biol.* **157** 1-7.

[132] Horwitz M.S.Z., D.K. Dube, and L.A. Loeb. (1989). Selection of new biological activities from random nucleotide sequences: evolutionary and practical considerations. *Genome* **31** 112-117.

[133] Tuerk, C. and L. Gold. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249** 505-510.

[134] Ellington A.D. and J.W. Szostak. (1990). *In vitro* selection of RNA molecules that bind to specific ligands. *Nature* **346** 818-822.

[135] Beaudry A.A. and G.F. Joyce. (1992). Directed evolution of an RNA enzyme. *Science* **257** 635-641.

[136] Bartel D.P. and J.W. Szostak. (1993). Isolation of New Ribozymes from a Large Pool of Random Sequences. *Science* **261** 1411-1418.

[137] Bauer, G.J., J.S.McCaskill, and A.Schwienhorst. (1989). Evolution im Laboratorium. *Nachr. Chem. Tech. Lab.* **37** 484-488.

[138] Wigner E. (1961). The logic of personal knowledge. In: Shils, E., ed. Essays presented to Michael Polanyi on his seventieth birthday 11th March 1961. Free Press: Glencoe (Ill.) .

[139] Cech T.R. (1986). RNA as an enzyme. *Sci. Am.* **255/5** 76-84.

[140] Guerrier-Takada C., K. Gardiner, T. Marsh, N. Pace, and S. Altman. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35** 849-857.

[141] Guerrier-Takada C. and A. Altman. (1984). Catalytic activity of an RNA molecule prepared by transcription *in vitro*. *Science* **223** 285-286.

[142] Symons R.H. (1992). Small catalytic RNAs. *Ann. Rev. Biochem.* **61** 641-671.

[143] Noller H.F. (1991). Ribosomal RNA and translation. *Ann. Rev. Biochem.* **60** 191-227.

[144] Noller H.F., V. Hoffarth, and L. Zimniak. (1992). Unusual resistence of peptidyl transferase to protein extraction procedures. *Science* **256** 1416-1419.

[145] Piccirilli J.A., T.S. McConnell, A.J. Zaug, H.F. Noller, and T.R. Cech. (1992). Aminoacyl esterase activity of the *tetrahymena*ribozyme. *Science* **256** 1420-1424.

[146] Gilbert W. (1986). The RNA World. *Nature* **319** 618.

[147] Joyce G.F. (1991). The rise and fall of the RNA world. *The New Biologist* **3** 399-407.

[148] Schuster P. (1981). Prebiotic Evolution. In: Gutfreund, H., ed. Biochemical Evolution. Cambridge Univ. Press: Cambridge (U.K.) pp.15-87.

[149] Mason S.F. (1991). Chemical evolution. Origin of the elements, molecules and living systems. Clarendon Press: Oxford (U.K.) .

[150] Orgel L.E. (1986). RNA catalysis and the origins of life. *J. Theor. Biol.* **123** 127-149.

[151] Joyce G.F. (1989). RNA evolution and the origins of life. *Nature* **338** 217-224.

10. Table of Contents