Fitness Landscapes Arising from the Sequence-Structure Maps of Biopolymers

Peter F. Stadler[a,b,*]

[a]Institut für Theoretische Chemie
Universität Wien, Vienna, Austria

[b]Santa Fe Institute, Santa Fe, NM

[*]Mailing Address:
Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
Phone: [43] 1 40480 665   Fax: [43] 1 40480 660
E-Mail: `studla@tbi.univie.ac.at`

Abstract

Fitness landscapes are an important concept in molecular evolution since evolutionary adaptation as well as *in vitro* selection of biomolecules can be viewed as a hill-climbing-like process. Global features of landscapes can be described by statistical measures such as correlation functions or the fraction of neutral (equally fit) neighbors. Simple spin-glass-like landscape models borrowed from statistical physics lend themselves to detailed mathematical analysis but lack several basic features of natural landscapes.
Biologically relevant landscape models are based on the assumption that genotypes give rise to phenotypes that are evaluated by their environment and hence determine the genotype's fitness. In the case of *in vitro* evolution of biopolymers the phenotypes are the three dimensional shapes of the molecules. A large degree of neutrality, giving rise to neutral networks and shape space covering, is a generic feature of RNA and polypeptide sequence-structure maps. These properties are inherited by the fitness landscapes independent of the details of the structure-to-fitness evaluation.
Neutrality qualitatively changes the dynamics of evolution. While rugged landscapes without neutral neighbors lead to localized populations, trapping in local optima, and the existence of a critical replication rate beyond which sequence information is lost, we find diffusion in sequence space and ever-lasting innovation of novel mutants on landscapes arising from RNA or protein folding.

Keywords

Fitness Landscapes — Molecular Evolution — RNA Secondary Structures — Biopolymer Folding — Graph Laplacian

1. Introduction

Since Sewall Wright's seminal paper [1] the notion of a *fitness landscape* underlying the dynamics of evolutionary optimization has proved to be one of the most powerful concepts in evolutionary theory. Implicit in this idea is a collection of genotypes arranged in an abstract metric space, with each genotype next to those other genotypes which can be reached by a single mutation, as well as a fitness value assigned to each genotype.

Such a construction is by no means restricted to biological evolution; Hamiltonians of disordered systems, such as spin-glasses [2, 3], and the cost functions of combinatorial optimization problems [4] have the same basic structure. A theory of landscapes, therefore, is based on three ingredients: we are given a *finite*, but very large set V of "configurations" and a "fitness function" $f : V \to \mathbb{R}$. The third ingredient is a notion of neighborhood between the configurations, which allows us to interpret V as the vertex set of a graph Γ. We will refer to Γ as the *configuration space* of the landscape f. The most prominent class of configuration spaces are the *sequence spaces* $\mathcal{Q}^n_\alpha$ consisting of all strings of length n composed from an alphabet with α letter. Two strings are neighbors of each other if they differ only in a single position. These graphs are known as Hamming graphs. The Hamming distance measures the number of positions in which two strings differ [5].

Conceptually, there is a close connection between (biological) *landscapes* and *potential energy surfaces* (PES) which constitute one of the most important issues of theoretical chemistry [6, 7]. As a consequence of the validity of the Born-Oppenheimer approximation, the PES provides the potential energy $U(\vec{R})$ of a molecule as a function of its nuclear geometry $\vec{R}$. PES are therefore defined on a high-dimensional *continuous* space and they are assumed to be smooth (usually twice continuously differentiable almost everywhere). The (global) analysis of PES thus makes extensive use of differential topology. The analysis of *discrete* landscapes, on the other hand, requires different techniques. For instance, the critical points of a PES, characterized by $\nabla U(\vec{R}) = 0$, have no obvious discrete counterpart.

It has been known since Eigen's [8, 9] pioneering work on the molecular quasispecies that the dynamics of evolutionary adaptation (optimization) on a landscape depends crucially on the detailed structure of the landscapes itself. Extensive computer simulations [10, 11] have made it very clear that a complete understanding of the dynamics is impossible without a thorough investigation of the underlying landscape [12]. Landscapes derived from well-known combinatorial optimization problems such as the Traveling Salesman Problem `TSP` [13], the Graph Bipartitioning Problem `GBP` [14], or the Graph Matching Problem `GMP` have been investigated in some detail, see [15] and the references therein. A detailed survey of a variety of model landscapes obtained by folding RNA molecules into their secondary structures has been performed during the last decade, see [16, 17, 18] and the references therein. While the use of (computationally simple) landscapes derived from spin-glasses or combinatorial optimization problems, or of the closely related Nk model [19] is certainly appealing, it is by no means clear that these models will capture the most salient features of biochemically relevant landscapes. Indeed, we shall show in this contribution that landscapes derived from folding biopolymers into their spatial structures are quite different from spin-glass-like landscapes.

One of the most important characteristic of a landscape is its *ruggedness*, a notion that is closely related to the hardness of the optimization problem for heuristic algorithms [20]. Three distinct approaches haven been proposed to measure and quantify ruggedness and to subsequently compare different landscapes. Sorkin [21], Eigen *et al.* [12] and Weinberger [22] used pair correlation functions. Kauffman and Levin [23] proposed adaptive walks, and Palmer [24] based his discussion on the number of meta-stable states (local optima). Of course one expects a close relationship between these different characterizations of ruggedness, which we shall discuss in some detail in section 2.

Mapping genotypes into fitness values is a core issue of evolutionary biology. It is commonly simplified by considering two separate steps:

$$\textbf{\textit{Genotype}} \quad \Longrightarrow \quad \textbf{\textit{Phenotype}} \quad \Longrightarrow \quad \textbf{\textit{Fitness}}$$

Genotype-phenotype mappings are generally too complicated to be analyzed by rigorous techniques. *In vitro* evolution of molecules, however, reduces this map to relations between sequences and structures of biopolymers. In section 3 we shall review the properties of sequence-structure maps of nucleic acids and proteins, and we shall see that the combined fitness landscapes inherit their most important properties from the underlying sequence-structure maps. The most important feature of all examples considered so far is *neutrality*: A very large number of sequences folds into the same structure[1]. Consequently, a large number of sequences have the same fitness. This high degree of neutrality distinguishes "biological" landscapes from the models borrowed from statistical mechanics. In section 4 we shall discuss the influence of neutrality on the dynamics of evolution.

2. Rugged Landscapes

The mathematical investigation of a landscape f on a graph Γ requires an algebraic description of the graph itself. The most straightforward encoding of Γ is the *adjacency matrix* $\mathbf{A}$ with entries $\mathbf{A}_{xy} = 1$ if the vertices x and y are connected by an edge, and $\mathbf{A}_{xy} = 0$ if x and y are not neighbors of each other. The degree matrix $\mathbf{D}$ of Γ is the diagonal matrix where $\mathbf{D}_{xx}$ is the number of neighbors of vertex x. All configuration spaces mentioned in this contribution are regular graphs, hence $\mathbf{D} = D\mathbf{I}$ where D is common degree of all vertices and $\mathbf{I}$ denotes the identity matrix. It is often more useful to use the *graph Laplacian* $-\Delta \stackrel{\text{def}}{=} \mathbf{D} - \mathbf{A}$. The graph Laplacian shares its most important properties with the familiar Laplacian differential operator: it is symmetric, non-negative definite, and singular

[1] In the puristic view of X-ray crystallography of biopolymers, sequence redundancy is non-existent: Small as they may be there are always differences in atomic coordinates that make structures unique. The crystallographic notion of structure, however, is vastly different from biochemical and evolutionary intuitions. Protein and RNA structures are often represented by wire diagrams. Phylogenetic conservation of structure is discussed, for example, by comparison of backbone foldings.

(the eigenvector $\mathbf{1} = (1, \ldots, 1)$ belongs to the eigenvalue $\Lambda_0 = 0$). There is also an analogue of Green's formula. For more details see [25, 15]. The graph Laplacian is central to the theory of electrical networks, see e.g. [26] and Kirchhoff's classical paper [27]. The formalism can be extended to hypergraphs derived from recombination [28, 29, 30].

A series expansion of a function in terms of a complete and orthonormal system of eigenfunctions of the Laplace operator is commonly termed Fourier expansion. We will adopt the same terminology here following [31]. Let $\{\varphi_i\}$ denote a complete orthonormal set of eigenvectors of $-\Delta$. We call the expansion

$$f(x) = \sum_{i=1}^{|V|} a_i \varphi_i(x) \tag{1}$$

a *Fourier expansion* of the landscape f. A non-flat landscape f is *elementary* if it is an eigenfunction of the graph Laplacian up to an additive constant, i.e., if and only if

$$\varphi(x) \stackrel{\text{def}}{=} f(x) - \frac{1}{|V|} \sum_{z \in V} f(z) \tag{2}$$

is an eigenfunction of $-\Delta$ with a non-zero eigenvalue $\Lambda > 0$. This definition is motivated by Lov Grover's observation [32] that the cost functions of a number of well-studied combinatorial optimization problems satisfies this condition for natural choices of move sets, see table 1.

Elementary landscapes play an important role because of their algebraic properties. It is easy to show that all local optima have fitness values above the average $\bar{f}$, and all local minima have fitness values below the average $\bar{f}$ [32]. The graph analogue of Courant's celebrated nodal domain theorem for Riemannian manifolds, see e.g. [33], was proved recently [34]: A *nodal domain* is a maximal connected subgraph of Γ on which f does not change sign. Suppose the eigenvalues of $-\Delta$ are labeled in ascending order

$$\Lambda_0 < \Lambda_1 \leq \Lambda_2 \leq \ldots \leq \Lambda_{k-1} \leq \Lambda_k \leq \Lambda_{k+1} \leq \ldots \leq \Lambda_{|V|-1} \,. \tag{3}$$

and repeated according to multiplicity. Let φ_k be *any* real valued eigenvector associated with the eigenvalue Λ_k. Courant's theorem then states that $k+1$ is an upper bound on the number of nodal domains of the eigenfunction φ_k.

The second-smallest eigenvalue Λ_1 of a graph and the corresponding eigenvectors have received some attention in algebraic graph theory. Kauffman [19] calls the

Table 1. Parameters of Elementary Landscapes.

Problem	Move Set	D	Λ	ℓ/n
NAES	Hamming	n	4	$1/4$
p-spin	Hamming	n	$2p$	$1/(2p)$
WP	Hamming	n	4	$1/4$
GC	Hamming	$(\alpha-1)n$	2α	$(1-1/\alpha)/2$
XY-Hamiltonian	Hamming	$(\alpha-1)n$	2α	$(1-1/\alpha)/2$
	cyclic	$2n$	$8\sin^2(\pi/\alpha)$	$1/[4\sin^2(\pi/\alpha)]$
GBP	Exchange	$n^2/4$	$2(n-1)$	$1/8{\cdot}n/(n-1)$
symmetric TSP	Transposition	$n(n-1)/2$	$2(n-1)$	$1/4$
	Inversions	$n(n-1)/2$	n	$(1-1/n)/4$
GMP	Transposition	$n(n-1)/2$	$2(n-1)$	$n/4$

The size of system n denotes the sequence length, the number of spins, or the number of cities in a traveling salesman problem.
The values K for NAES (Non-All-Equal-Satisfiability), WP (Weight Partition), GC (Graph Coloring with α colors), GBP (Graph Bipartitioning), and TSP (Traveling Salesman Problem) are taken from [32]. The value of Λ for the GMP (Graph Matching Problem) is derived in [15]. The values of λ for the GBP and the GMP problem are taken from [35] and [36], respectively. The configuration space of the XY-Hamiltonian $\sum_{i<j} J_{ij}\cos(\frac{2\pi}{\alpha}(x_i - x_j))$ is either a sequence space with α letters (denoting) the spin positions, or the direct sum of n cycles, if one assumes that spin my move only by $\pm 2\pi/\alpha$ [37].

corresponding landscapes *Fujijama*, because they have only a single mountain massive (positive nodal domain). On sequence spaces these cost functions are always additive, i.e., the fitness is the sum of contributions from the individual positions (monomers). Recently Fujijama landscapes have been discussed as models of binding energy landscapes of oligonucleotides [38]. In contrast, almost all the landscapes listed in table 1 (with the exception of the p-spin models for $p > 2$), belong to the third-smallest eigenvalue and hence to the simplest class of truely rugged landscapes.

Two types of *correlation functions* have been investigated as a means of quantifying the ruggedness of a landscape. Eigen and co-workers [12] introduced $\rho(d)$ which measures the pair correlation as a function of the distance between the vertices of Γ. Weinberger [22] used the autocorrelation function $r(s)$ of the "time series" $\{f(x_0), f(x_1), \ldots\}$ generated by a simple random walk [39] on Γ in order to measure properties of f. The relationship between $r(s)$ and $\rho(d)$ is discussed in [22, 40].

The correlation function $r(s)$ is intimately related to the Fourier series expansion of the landscape [15]. Elementary landscapes belonging to the eigenvalue Λ_p have exponential autocorrelation functions of the form $r(s) = (1 - \Lambda_p/D)^s$. For any landscape holds

$$r(s) = \sum_{p \neq 0} B_p (1 - \Lambda_p/D)^s \,. \tag{4}$$

The *amplitudes* B_p are determined by the Fourier coefficients a_k in equ.(1):

$$B_p = \sum_{k \in I_p} |a_k|^2 \Big/ \sum_{k \neq 0} |a_k|^2 \ \geq 0 \,, \tag{5}$$

where I_p denotes the set of the indices j for which $-\Delta\varphi_j = \Lambda_p \varphi_j$. The crucial information about a landscape is therefore contained in the eigenvalues Λ_p of the graph Laplacian, which determine the ruggedness of a component, and in the amplitudes B_p, which determine the relative importance of the different modes. A particularly useful measure for the ruggedness of a landscape is the correlation length

$$\ell \stackrel{\text{def}}{=} \sum_{s=0}^{\infty} r(s) = D \sum_{p \neq 0} \frac{B_p}{\Lambda_p} \tag{6}$$

[22, 40, 41, 42]. This quantity can be estimated rather easily in (computer) experiments. For an elementary landscape we have $\ell = D/\Lambda$.

Most landscape models contain a stochastic element in their definition: a particular instance is generated by assigning a (usually) large number of parameters at random. Such models are called *random fields* [43]. A typical example is the Sherrington-Kirkpatrick Hamiltonian [44], $f(x) \stackrel{\text{def}}{=} \sum_{i<j} J_{ij} x_i x_j$, where $x = (x_1, \ldots, x_n)$ denotes a configurations of spins $x_i = \pm 1$, and the "coupling constants" are identically and independently distributed (i.i.d.) Gaussian random variables. We shall write $\mathcal{E}[\,.\,]$ for the *average over the disorder*, i.e., the random variables in the landscape model. In the SK model this amounts to integrating over the Gaussian distributions of the interaction coefficients J_{ij}.

A fairly general algebraic theory of isotropy is laid out in [45]. A random field on a graph Γ is *isotropic* if its covariance matrix

$$\mathbf{C}_{xy} = \mathcal{E}[f(x)f(y)] - \mathcal{E}[f(x)]\mathcal{E}[f(y)] \tag{7}$$

is invariant under all automorphisms of Γ. The following proposition characterizes isotropy in terms of the Fourier coefficients:

Proposition. A random field on a sequence space is *isotropic* if and only if its Fourier coefficients $\{a_k\}$ fulfill

(i) $\mathcal{E}[a_k] = 0$ for all $k \neq 0$;
(ii) $\mathcal{E}[a_k a_l^*] = \delta_{kl}\mathcal{E}[|a_k|^2]$;
(iii) $\mathcal{E}[|a_j|^2] = \mathcal{E}[|a_k|^2] = \beta_p$ whenever the corresponding eigenfunctions φ_j and φ_k belong to the same eigenvalue Λ_p of the graph Laplacian.

This observation suggests to interpret isotropy as a *maximum entropy* like condition: Given the parameters β_p, the "most random" choice of coupling constants are Gaussian random variables fulfilling (i) through (iii). Derrida's p-spin models [46], for instance, are the maximum entropy models with the single constraint that only one order of interaction contributes to the Hamiltonian, and the random energy model [47] can be regarded as the maximum entropy model subject to the constraint that the constants β_p are all equal [45].

Palmer [24] used the existence of a large number of local optima to define ruggedness. We say that $x \in V$ is a local minimum of the landscape f if $f(x) \leq f(y)$ for all neighbors y of x. The use of $\leq$ instead of $<$ is conventional [48, 49]. The number $\mathcal{N}$ of local optima of a landscape is much harder to determine than its autocorrelation function $r(s)$ or its correlation length ℓ. A heuristic argument linking local optima and correlation measures runs as follows: For a typical elementary landscape we expect that the correlation length ℓ gives a good description of its structure because the landscape does not have any other distinctive features. By construction ℓ determines the size of the mountains and valleys. As there are many directions available at each configuration we expect there are only very few metastable states besides the summit of each of these ℓ-sized mountains – almost all of the configurations will be saddle points with at least a few superior neighbors. We measure ℓ along a random walk but the radius $R(\ell)$ of a mountain is more conveniently described in terms of the distance between vertices on Γ. Here $R(\ell)$ is the average distance that is reached by the random walk in ℓ steps. With the notation $B(R)$ for the number of vertices contained in a ball of radius R in Γ we obtain the estimate $\mathcal{E}[\mathcal{N}] \approx |V|/B(R(\ell))$ local optima. There is a fair amount of computational evidence for the correlation length conjecture [37] in isotropic and nearly isotropic elementary landscapes. In addition, one obtains reasonably good estimates for Kauffman's Nk landscapes [50]. A few counter-examples are known as well; all of them strongly violate the maximum entropy assumption.

3. Structure-Based Landscapes

Mapping genotypes into fitness values is a core issue of evolutionary biology. It is commonly simplified by partitioning the task in two steps, namely formation of the phenotype from the genotype and subsequent evaluation of the phenotype, see figure 1. *In vitro* evolution of biomolecules, however, reduces this map to relations between polynucleotide sequences and biopolymer structures and functions.

RNA is a particularly fruitful system for the computational (bio)chemist because the structure prediction problem can be solved efficiently at least at the approximate level of secondary structures: The total energy of folding an RNA molecule into its secondary structure can be approximated by additive contributions for stacking of Watson-Crick (**GC** and **AU**) and **GU** base pairs and by destabilizing contributions for loops. The secondary structure is precisely the list of these base pairs; it can be represented by a planar graph without knots or pseudo-knots[2].

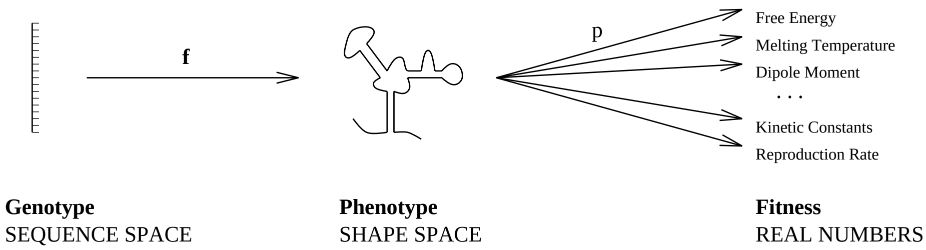


Figure 1: Landscapes based on genotype mappings can be viewed as compositions $p(f(g))$, where f : Sequence Space $\to$ Shape Space represents folding and p : Shape Space $\to \mathbb{R}$ encodes the evaluation of the structure by the environment.

Experimental energy parameters are available for the individual contributions as functions of type (stacked pair, interior loop, bulge, multi-stem loop), of size, of the type of the delimiting base pairs, and partly of the sequence of the unpaired subsequences, see e.g. [51]. As a consequence of the additivity of the energy contributions, the minimum energy of an RNA sequence can be calculated recursively by dynamic programming [52, 53]. An efficient implementation of this algorithm is part of the `Vienna RNA Package` [54] which is freely available from

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

www.tbi.univie.ac.at. Some statistical properties of RNA secondary structures were shown to depend very little on choices of algorithms and parameter sets [55].

It is possible to derive exact recursions enumerating secondary structure graphs [52, 56]. From these recursions one obtains, for instance, an asymptotic expression for the numbers S_n of (acceptable) structures that can be formed by sequences of chain length n:

$$S_n \;\approx\; 1.4848 \times n^{3/2}\,(1.8488)^n \;. \tag{8}$$

Equ.(8) is based on two assumptions: (i) the minimum stack length is two base pairs (i.e., isolated base pairs are excluded) and (ii) the minimal size of hairpin loops is three. The number of acceptable structures with pseudo-knots increases asymptotically with $S_n \propto 2.35^n$ [57]. In contrast, there are 4^n possible RNA sequences composed from the natural **AUGC** alphabet; thus many sequences must fold into the same structure.

Secondary structures are properly grouped into two classes, common ones and rare ones. A structure ψ is *common* if it is formed by more sequences than the average structure. Data from both large samples of long sequences ($n \gg 30$) [58, 59] and from exhaustive folding of all short sequences [60, 61] support two important observations: (i) the common structures represent only a small fraction of all structures and this fraction decreases with increasing chain length; (ii) the fraction of sequences folding into common structures increases with chain length and approaches 100% in the limit of long chains. Thus, for sufficiently long chains almost all RNA sequences fold into a small fraction of the secondary structures. The effective ratio of sequences to structures is even larger than computed from equ.(8) since only common structures play a role in natural evolution and in evolutionary biotechnology [59]. RNA and proteins, despite their different chemistry, apparently share fundamental properties of their sequence-structure maps: the repertoire of stable native folds seems to be highly restricted or even vanishingly small [62].

Naturally, we ask how sequences folding into the same (common) secondary structure are distributed in sequence space. We call the set $S(\psi)$ of all sequences (genotypes) folding into phenotype ψ the *neutral set* of ψ[3]. The shape or topology of neutral sets has important implications for the evolution of both nucleic acids and proteins and for *de novo* design: For example, it has been frequently observed that seemingly unrelated protein sequences have essentially the same fold [63, 64, 65]. Similarly, the genomic sequences of closely related RNA viruses show a large

[3] A mathematician would call S the pre-image of ψ w.r.t. the folding map f.

degree of sequence variation while sharing many conserved features in their secondary structures [66, 67]. Another well known example is the clover leaf secondary structure of tRNAs: The sequences of different t-RNA's have little sequence homology [68] but nevertheless fold into the same secondary structure motif. Whether similar structures with distant sequences may have originated from a common ancestor, or whether they must be the result of convergent evolution, depends on the geometry of the neutral sets $S(\psi)$ in sequence space.

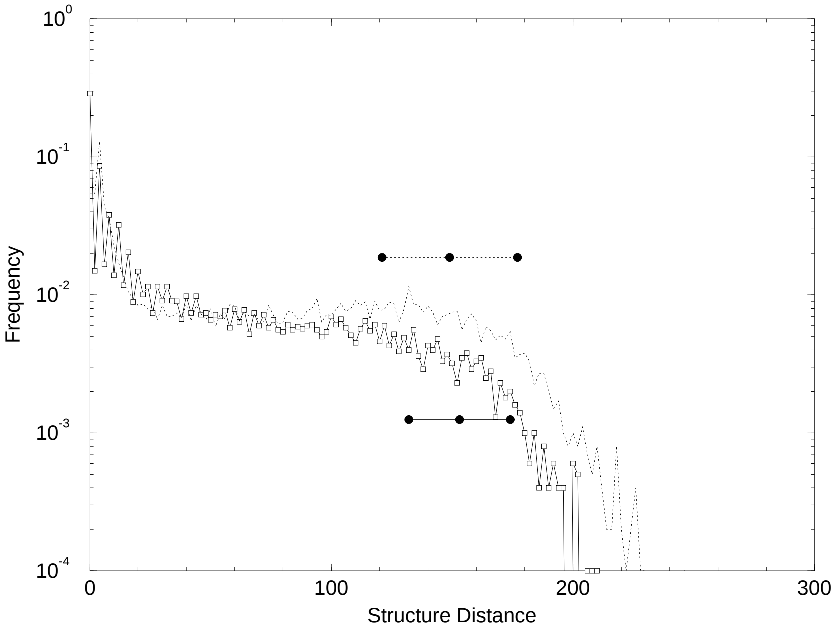


Figure 2: Distribution of structure distances between RNA sequences differing by a single point mutations, $n = 200$. Full line: natural **GCAU** alphabet, dotted line: **GC** alphabet. About 30% of the sequence pairs fold into the same structure. This high degree of neutrality implies the existence of connected neutral networks. On the other hand, a substantial fraction of point mutations leads to structure distances comparable to the structure distances between random sequences (mean and one standard deviation are indicated by circles). The structure distance is defined as edit distance on the tree representations of secondary structure graphs, see [69, 54] for details.

Inverse folding can be used to determine the sequences that fold into a given structure. Naturally, a sequence x can fold into a given secondary structure ψ only if each pair of sequences positions that is paired in ψ is realized by one of the six possible base pairs. The set of all such sequences forms $C(\psi)$, the set of *compatible* sequences. Of course we have $S(\psi) \subset C(\psi)$. For RNA secondary structures an efficient inverse folding algorithm is available [54]. It was used to show that sequences folding into the same structure are (almost) randomly distributed in the space $C(\psi)$ of compatible sequences. A similar result was obtained for "protein space" [70] using so-called knowledge-based potentials of mean force [71, 72, 73] for deciding whether a given sequence x folds into a native protein fold ψ. On the other hand, it was noticed already in early work on RNA secondary structures [10] that a substantial fraction of point mutations are neutral, i.e., that many sequences differing only in a single position fold into the same secondary structure, see figure 2.

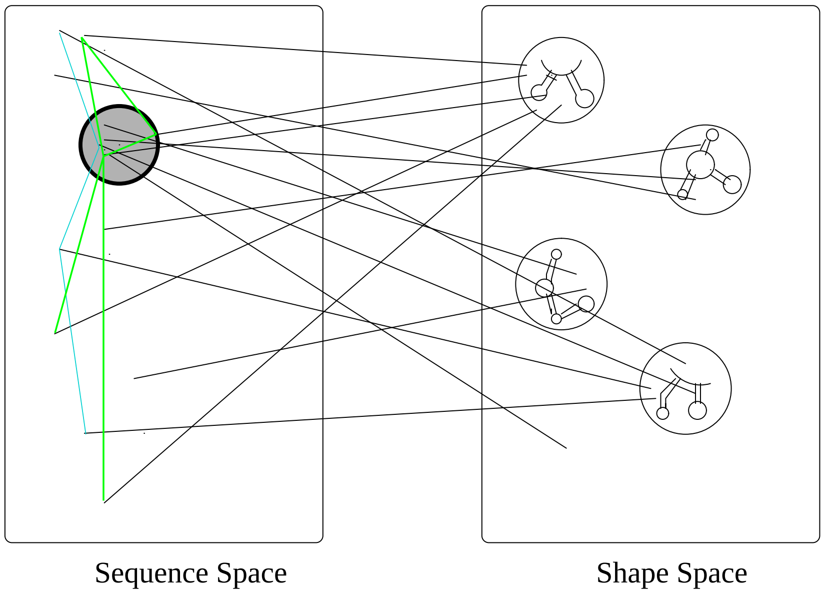


Figure 3: Sequence-Structure Map of Biopolymers.
Sequences folding into the same structure lie on a connected network in sequence space. All structures are formed from some of the sequences contained in a small ball around an arbitrary reference point in sequence space.

Three approaches have been applied so far to study the topology of neutral sets: a mathematical model of genotype-phenotype mapping based on random graph theory [74], extensive sample statistics [58] and exhaustive folding of all sequences with given chain length n [61]. The mathematical model assumes that sequences forming the same structure are distributed randomly using the fraction λ of neutral neighbors as (the only) input parameter. If λ is large enough this model makes two rather surprising predictions [74, 75]:

(1) The connectivity of networks changes drastically when λ passes the threshold value:

$$\lambda_{cr}(\alpha) = 1 - \sqrt[\alpha-1]{\frac{1}{\alpha}}\,, \tag{9}$$

where α is the size of the alphabet. Neutral sets consist of a single component that span the sequence space if $\lambda > \lambda_{cr}$ and below threshold, $\lambda < \lambda_{cr}$, the network is partitioned into a large number of components, in general, a giant component and many small ones. In the first case we refer to $S(\psi)$ as the *neutral network* of ψ. For RNA it is necessary to split the random graph into two factors corresponding to unpaired bases and base pairs and to use a different value of λ for each factor. Each of these two parameters is much larger than the critical value for common RNA secondary structures, hence the neutral sets $S(\psi)$ form form connected neutral networks within the sets $C(\psi)$ of compatible sequences [74]. The situation appears to be similar for proteins [70].

(2) There is *shape space covering*, that is, in a moderate size ball centered at any position in sequence space there is a sequence x that folds into any prescribed secondary structure ψ. The radius of such a sphere, called the *covering radius* r_{cov}, can be estimated from simple probability arguments [59]

$$r_{cov} \quad \approx \quad \min\left\{h \;\middle|\; B(h) \geq S_n\right\}, \tag{10}$$

with $B(h)$ being the number of sequences contained in a ball of radius h. The covering radius is much smaller than the radius n of sequence space. The covering sphere represents only a small connected subset of all sequences but contains, nevertheless, all common structures and forms an evolutionarily representative part of shape space.

Figure 3 is a sketch of a typical sequence-structure map. The existence of extensive neutral networks meets a claim raised by Maynard-Smith [76] for protein spaces that are suitable for efficient evolution. The evolutionary implications of neutral networks are explored in detail in [77, 78] and will be reviewed in the following section. Empirical evidence for a large degree of *functional* neutrality in protein space was presented recently by Wain-Hobson and co-workers [79].

The ruggedness of sequence-structure maps can be computed in terms of the generalization

$$r(s) = 1 - \frac{\langle \mathcal{D}^2(f(x_t), f(x_{t+s}))\rangle}{\langle \mathcal{D}^2\rangle}, \tag{11}$$

of the random walk correlation function $r(s)$ see [41]. Here $\mathcal{D}(\psi, \psi')$ is a distance measure in shape space[4], and $\langle \mathcal{D}^2\rangle$ is the average value over a sample of random sequences. RNA secondary structure correlation functions are surprisingly rugged

[4]One may use the trivial structure distance $\mathcal{D}(\psi, \psi') = 1 \iff \psi \neq \psi'$ or a more elaborate one such as the RNA tree-edit distance [69] without significantly affecting the results.

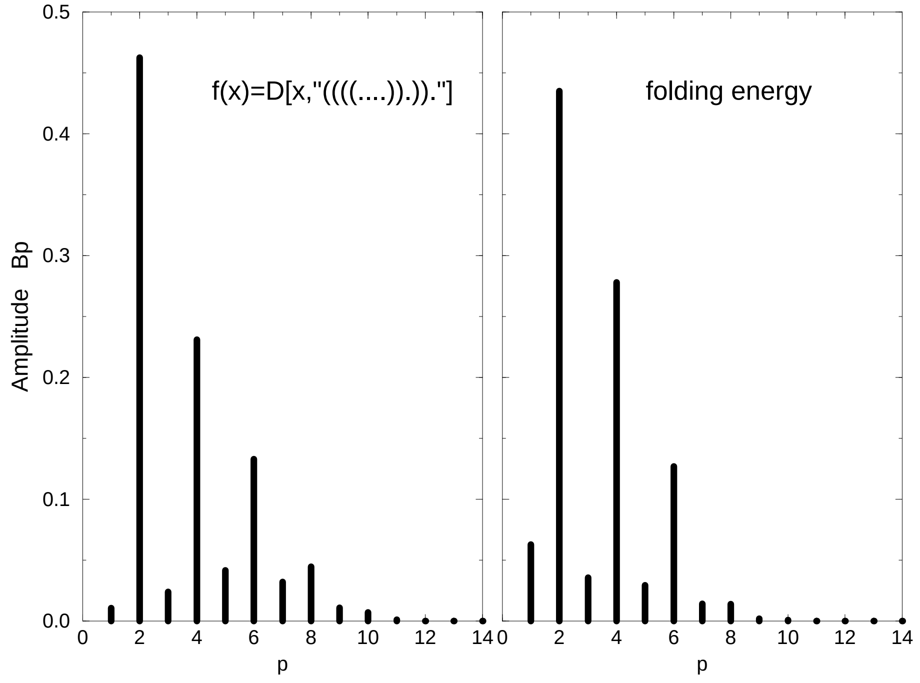


Figure 4: Amplitude spectrum of two RNA landscape with $n = 14$. The amplitudes B_p are computed using FFT and equ.(5). L.h.s.: The fitness function is defined as $f(x) = D(x, T)$ where the target structure $T =$ `'((((..)).)).'`, and D denotes the tree edit distance [69]. R.h.s.: The fitness equals the energy of folding sequence x into its secondary structure.
The amplitude spectrum of these two landscapes is surprisingly similar despite their quite different definitions. The fact that odd interaction orders play only a minor role reflects the fact that base pairing and stacking of base pairs, which involves always an even number of nucleotides, is the dominating stabilizing energy contribution. The correlation lengths are $\ell = 2.454$ and $\ell = 2.752$, respectively.

despite the high degree of neutrality in RNA as a consequence of shape space covering: a substantial fraction of all mutations lead to very different structures and hence to high a large value of $\mathcal{D}^2(f(x_t), f(x_{t+s}))$ even for $s = 1$. The structure correlation length of RNA secondary structures, for instance, is $\ell_{\rm str} \approx 0.0524n$, or only about one fifth of the correlation length a typical spin-glass model [16].

Landscapes based on sequence-structure maps of course inherit their ruggedness even if the map from structures to fitness values is smooth or even linear, since shape space covering implies that a substantial fraction of point mutations lead to unrelated structures. On the other hand, a completely random assignment of fitness values to structures cannot undo the correlation introduced by neutrality: In this case the expected correlation function of the fitness landscape equals the correlation function (11) of the sequence-structure map computed from the trivial structure distance. As shown in [74], we have $r(s) \approx \bar{\lambda}(s)$, the probability of finding a neutral structure after s steps of the random walk in this case. The fundamental properties of structure-based landscapes are therefore properly described by the underlying sequence-structure map.

Not surprisingly, structure-based landscapes are far from being elementary, see figure 4 for two examples. Their amplitude spectra show a rather broad distribution of contributing interaction orders and oftentimes a distinct pattern that can be explained in terms of the biophysical properties of the underlying molecules. Similar features were described recently for landscapes arising from the synchronization problem of cellular automata [80].

4. Landscape Structure and the Dynamics of Evolution

Simplifying the detailed mechanisms of replication and mutation one may represent the dynamics of evolution by a reaction-diffusion equation of the form [81, 82, 83]

$$\frac{\partial}{\partial t}\phi(x,t) = \delta\,\Delta\phi(x,t) + \phi(x,t)\Big(F(x,\vec{\phi}) - \Phi(t)\Big)\,, \tag{12}$$

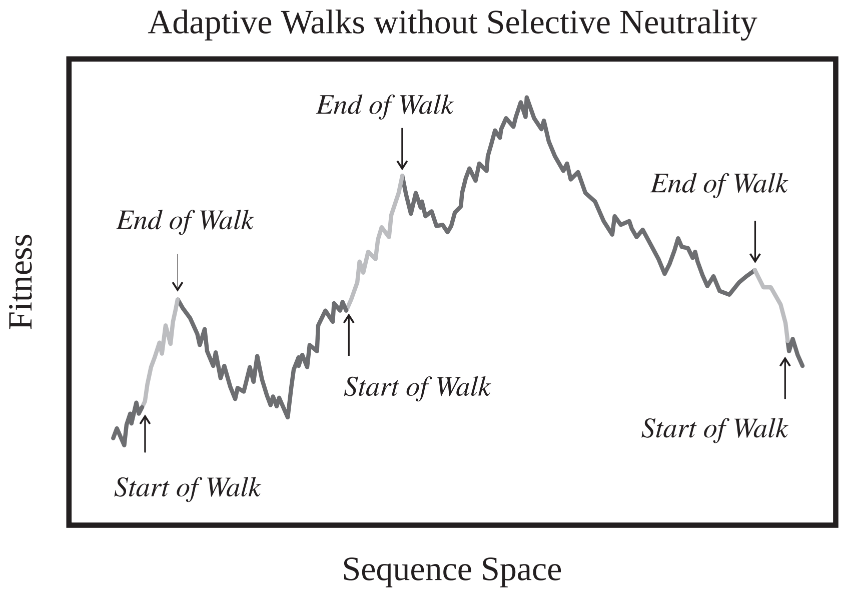
where $\phi(x,t)$ denotes the fraction of genotypes x at time t and $\Phi(t) = \sum_x F(x,\vec{\phi})$ is an unspecific dilution term ensuring conservation of probability. In general $F(x,\vec{\phi})$ will be a non-linear function of the genotype frequencies describing the interactions between different species as well as their autonomous growth [84]. Within the context of this contribution $F(x,\vec{\phi}) = f(x)$, the fitness landscape. The diffusion constant is $\delta = (1-Q)\max_x F(x)/D$, where Q is the probability of correct replication. In terms of the more widely used single-digit mutation rate p we have $Q = (1-p)^n \approx 1 - np + \mathcal{O}(p^2)$, and hence $\delta \approx p\,F_{\max}/(\alpha-1)$ on a sequence space with α letters. While equ.(12) is not suitable for a detailed quantitative prediction of a particular model, it is a valuable heuristic for explaining some of the most important effects. One should keep in mind, however, that equ.(12) is a mean field equation that does not correctly describe some important effects even in the limit of large populations (see [85] for an instructive example).

Evolutionary dynamics on rugged landscapes without neutrality, such as the spin-glass like models discussed in section 2, are considered for instance in [8, 12, 82]. For small mutation rates p a population is likely to get stuck in local optima for very long times. Populations form localized quasi-species around a “master sequence”. There is a critical mutation rate p_{et} at which diffusion outweighs selection and the population begins to drift in sequence space – the genetic information is lost [8, 12]. As an order of magnitude estimate one finds $p_{et} \approx \sigma/n$ where the "superiority" σ is a measure of the fitness advantage of the master sequence.

On a flat fitness landscape, $f(x) = 1$ for all $x \in V$, the selection term disappears and we are left with a pure diffusion equation. A stochastic description can be found in [86]. The situation on landscapes with a large degree of neutrality is much closer to the flat landscape than a non-neutral rugged one, despite the fact that $r(s)$ may decay very rapidly. There is no stationary master species surrounded by a mutant cloud, since Eigen's superiority parameter σ is so small in the presence of a large number of neutral mutants that sensible values of p exceed the (genotypic) errortheshold by many orders of magnitude. For small values of p the neutral network of the fittest *structure*, $S(\psi)$, dominates the dynamics. Populations migrate by a diffusion-like mechanism [86, 77] on $S(\psi)$ just like on a flat landscape with the single modification that the effective diffusion constant is smaller by the factor λ, the fraction of neutral mutations.

Random drift is continued until the population reaches an area in sequence space where some fitness values are higher than that of the currently predominating neutral network. Then a period of Darwinian evolution sets in, leading to the selection of the locally fittest structure. Evolutionary adaptation thus appears as a stepwise process: phases of increasing mean fitness (transitions between different structures) are interrupted by periods of apparent stagnation with mean fitness values fluctuating around a constant (diffusion on a neutral network) [77], figure 5. When the fittest structure is common its neutral network extends through the entire sequence space allowing the population to eventually find the global fitness optimum. A population is not a single localized quasi-species in sequence space [12], but rather a collection of different quasi-species since population splits into well separated clusters [77] on a single neutral network. Each cluster undergoes independent diffusion, while all share the same dominant phenotype. It is not surprising hence that there are abundant examples of both RNA and protein structures that have been conserved over evolutionary time scales while the underlying sequences have lost (almost) all homology.

For larger mutation rates p the diffusion term in equ.(12) dominates the dynamics. Assuming that all sequences $x \notin S(\psi)$ have fitness g and $f(x) = f$ for $x \in S(\psi)$ we may compute the mean field time evolution of $\theta(t) = \sum_{x\in S(\psi)} \phi(x,t)$. Substituting this into equ.(12) we find that the diffusion term yields approximately $\delta(1-\lambda)\theta(t)$, accounting for the fraction $1-\lambda$ of offsprings that are not members of the neutral network $S(\psi)$. The replication term becomes $\theta(t)\,[f - \theta(t)f - (1-\theta(t))g]$. Hence $\theta(t)$, the fraction of sequences folding into the dominating phenotype, approaches



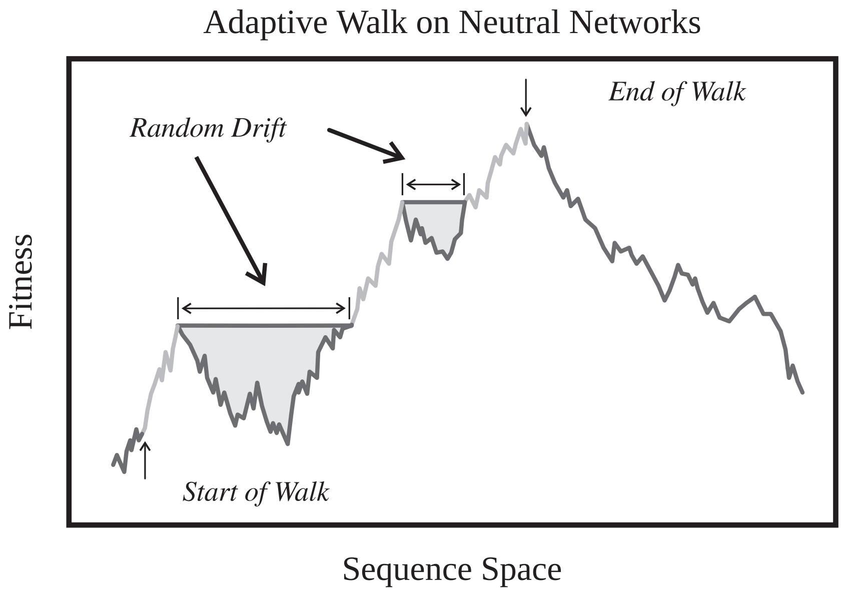


Figure 5: The role of neutral networks in evolution [87].
Optimization occurs through adaptive walks and random drift. Adaptive walks allow to choose the next step arbitrarily from all directions where fitness is (locally) non-decreasing. Populations can bridge over narrow valleys with widths of a few point mutations. In the absence of selective neutrality (spin-glass-like landscape, above) they are, however, unable to span larger Hamming distances and thus will approach only the next major fitness peak. Populations on rugged landscapes with extended neutral networks evolve along the networks by a combination of adaptive walks and random drift at constant fitness (below). In this manner, populations bridge over large valleys and may eventually reach the global maximum of the fitness landscape.

a stationary value $\theta = 1 - (1-\lambda)n\rho\sigma^*$, where $\sigma^* = (f-g)/f$ may be interpreted as "superiority" of the structure ψ. A crude estimate for the *phenotypic* error threshold, at which the dominating phenotype is lost, is obtained by setting $\theta = 0$:

$$p_{\text{phen.et.}} \approx \frac{1}{1-\lambda}\frac{\sigma^*}{n} \approx \frac{\sigma^*}{n}(1+\lambda) \tag{13}$$

A more careful derivation can be found in [88]. It shows that there is critical value $\lambda = g/f$ above which all error rates can be tolerated without loosing phenotype. A much more elaborate computation of the phenotypic error threshold can be found in [89]. The crude estimate (13) matches the available simulation results within a factor 3. Note that equ.(13) reduces to the estimate of Eigen's sequence error-threshold in the limit $\lambda \to 0$: this is sensible: an isolated sequence with fitness $f > g$ sustains a localized population for small enough mutation rates.

Diffusion in sequence space, the existence of phenotypic error threshold, and a close connection [77] with Kimura's neutral theory [81] which we have not discussed here, are consequences of the existence of neutral networks. Shape space covering implies a constant rate of innovation [78]: While diffusing along a neutral network, a population constantly produces non-neutral mutants folding into different structures. Shape space covering implies that almost all structures can be found somewhere near the current neutral network. Hence the population keeps discovering structures that it has never encountered before at a constant rate. When a superior structure is produced, Darwinian selection becomes the dominating effect and the population "jumps" onto the neutral network of the novel structure while the old network is abandoned. Figures 5 sketches the difference between evolutionary adaptation on spin-glass-like landscapes and on the highly neutral landscapes arising from biopolymer structures.

Neutral evolution, arising as a consequence of the high degree of neutrality observed in genotype-phenotype mappings of biopolymers, therefore, is not a dispensable addendum to evolutionary theory (as it has often been suggested). On the contrary, neutral networks, provide a powerful mechanism through which evolution can become truely efficient.

Acknowlegements

Discussions with Peter Schuster and Ivo Hofacker are gratefully acknowleged. Special thanks to Ivo Hofacker and Wim Hordijk for the data shown in figure 2 and part of figure 4, respectively.

References

[1] S. Wright. The roles of mutation, inbreeding, crossbreeeding and selection in evolution. In D. F. Jones, editor, *Int. Proceedings of the Sixth International Congress on Genetics*, volume 1, pages 356–366, 1932.

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

[3] M. Mézard, G. Parisi, and M. Virasoro. *Spin Glass Theory and Beyond.* World Scientific, Singapore, 1987.

[4] M. Garey and D. Johnson. *Computers and Intractability. A Guide to the Theory of $\mathcal{NP}$ Completeness.* Freeman, San Francisco, 1979.

[5] R. W. Hamming. Error detecting and error correcting codes. *Bell Syst.Tech.J.*, 29:147–160, 1950.

[6] P. G. Mezey. *Potential Energy Hypersurfaces.* Elsevier, Amsterdam, 1987.

[7] D. Heidrich, W. Kliesch, and W. Quapp. *Properties of Chemically Interesting Potential Energy Surfaces*, volume 56 of *Lecture Notes in Chemistry.* Springer-Verlag, Berlin, 1991.

[8] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften*, 10:465–523, 1971.

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

[10] W. Fontana and P. Schuster. A computer model of evolutionary optimization. *Biophysical Chemistry*, 26:123–147, 1987.

[11] W. Fontana, W. Schnabl, and P. Schuster. Physical aspects of evolutionary optimization and adaption. *Physical Review A*, 40:3301–3321, 1989.

[12] M. Eigen, J. McCaskill, and P. Schuster. The molecular Quasispecies. *Adv. Chem. Phys.*, 75:149 – 263, 1989.

[13] E. L. Lawler, J. K. Lenstra, A. H. G. R. Kan, and D. B. Shmoys. *The Traveling Salesman Problem. A Guided Tour of Combinatorial Optimization.* John Wiley & Sons, 1985.

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

[15] P. F. Stadler. Landscapes and their correlation functions. *J. Math. Chem.*, 20:1–45, 1996.

[16] P. Schuster and P. F. Stadler. Landscapes: Complex optimization problems and biopolymer structures. *Computers Chem.*, 18:295–314, 1994.

[17] P. Schuster, P. F. Stadler, and A. Renner. RNA Structure and folding. From conventional to new issues in structure predictions. *Curr. Opinion Struct. Biol.*, 7, 1997. 229-235.

[18] P. Schuster and P. F. Stadler. Sequence redundancy in biopolymers: A study on RNA and protein structures. In G. Myers, editor, *Viral Regulatory Structures*, volume XXVIII of *Santa Fe Institute Studies in the Sciences of Complexity*. Addison-Wesley, Reading MA, 1997. in press, Santa Fe Institute Preprint 97-07-67.

[19] S. Kauffman. *The Origin of Order.* Oxford University Press, New York, Oxford, 1993.

[20] B. Manderick, M. de Weger, and P. Spiessen. The genetic algorithm and the structure of the fitness landscape. In R. K. Belew and L. B. Booker, editors, *Proceedings of the 4th International Conference on Genetic Algorithms.* Morgan Kaufmann Inc., 1991.

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

[22] E. D. Weinberger. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biol. Cybern.*, 63:325–336, 1990.

[23] S. A. Kauffman and S. Levin. Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.*, 128:11, 1987.

[24] R. Palmer. Optimization on rugged landscapes. In A. S. Perelson and S. A. Kauffman, editors, *Molecular Evolution on Rugged Landscapes: Proteins, RNA, and the Immune System*, pages 3–25. Addison Wesley, Redwood City, CA, 1991.

[25] B. Mohar. The laplacian spectrum of graphs. In Y. Alavi, G. Chartrand, O. Ollermann, and A. Schwenk, editors, *Graph Theory, Combinatorics, and Applications*, pages 871–897, New York, 1991. John Wiley & Sons.

[26] P. M. Soardi. *Potential Theory on Infinite Networks*, volume 1590 of *Lecture Notes in Mathematics.* Springer-Verlag, Berlin, 1994.

[27] G. Kirchhoff. Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der lineare Verteilung galvanischer Ströme geführt wird. *Ann. Phys. Chem.*, 72:487–508, 1847.

[28] P. Gitchoff and G. P. Wagner. Recombination induced hypergraphs: A new approach to mutation-recombination isomorphism. *Complexity*, 2:37–43, 1996.

[29] P. F. Stadler and G. P. Wagner. The algebraic theory of recombination spaces. *Evol. Comp.*, 1997. in press, Santa Fe Institute Preprint 96-07-046.

[30] G. P. Wagner and P. F. Stadler. Complex adaptations and the structure of recombination spaces. In ?, editor, *Proceedings of the Conference on Semi-Groups and Algebraic Engineering*, University of Aizu, Japan, 1997. ? in press, Santa Fe Institute Preprint 97-03-029.

[31] E. D. Weinberger. Local properties of Kauffman's N-k model: A tunably rugged energy landscape. *Phys. Rev. A*, 44:6399–6413, 1991.

[32] L. Grover. Local search and the local structure of NP-complete problems. *Oper.Res.Lett.*, 12:235–243, 1992.

[33] I. Chavel. *Eigenvalues in Riemannian Geometry.* Academic Press, Orlando Fl., 1984.

[34] Y. Colin De Verdiere. Multiplicites des valeurs prores laplaciens discrets et laplaciens continus. *Rend. mat. appl.*, 13:433–460, 1993.

[35] P. F. Stadler and R. Happel. Correlation structure of the landscape of the graph-bipartitioning-problem. *J. Phys. A.: Math. Gen.*, 25:3103–3110, 1992.

[36] P. F. Stadler. Correlation in landscapes of combinatorial optimization problems. *Europhys. Lett.*, 20:479–482, 1992.

[37] R. García-Pelayo and P. F. Stadler. Correlation length, isotropy, and metastable states. *Physica D*, 107:240–254, 1997.

[38] T. Aita and Y. Husimi. Fitness spectrum among the mutants of mt. fuji-type fitness landscapes. *J. Theor. Biol.*, 182:469–485, 1996.

[39] F. Spitzer. *Principles of Random Walks.* Springer-Verlag, New York, 1976.

[40] W. Fontana, T. Griesmacher, W. Schnabl, P. Stadler, and P. Schuster. Statistics of landscapes based on free energies, replication and degredation rate constants of RNA secondary structures. *Monatsh. Chemie*, 122:795–819, 1991.

[41] W. Fontana, P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, and P. Schuster. RNA folding and combinatory landscapes. *Phys. Rev. E*, 47:2083 – 2099, 1993.

[42] R. Happel and P. F. Stadler. Canonical approximation of landscapes. *Complexity*, 2:53–58, 1996.

[43] J. Besag. Spatial interactions and the statistical analysis of lattice systems. *Amer. Math. Monthly*, 81:192–236, 1974.

[44] D. Sherrington and S. Kirkpatrick. Solvable model of a spin-glass. *Physical Review Letters*, 35(26):1792 – 1795, 1975.

[45] P. F. Stadler and R. Happel. Random field models for fitness landscapes. *J. Math. Biol.*, 1996. in press, Santa Fe Institute preprint 95-07-069.

[46] B. Derrida. Random energy model: Limit of a family of disordered models. *Phys.Rev.Lett.*, 45:79–82, 1980.

[47] B. Derrida. The random energy model. *Phys.Rep.*, 67:29–35, 1980.

[48] W. Kern. On the depth of combinatorial optimization problems. *Discr. Appl. Math.*, 43:115–129, 1993.

[49] J. Ryan. The depth and width of local minima in discrete solution spaces. *Discr. Appl. Math.*, 56:75–82, 1995.

[50] C. A. Macken and P. F. Stadler. Evolution on fitness landscapes. In L. Nadel and D. L. Stein, editors, *1993 Lectures in Complex Systems*, volume VI of *SFI Studies in the Sciences of Complexity*, pages 43–86. Addison-Wesley, Reading MA, 1995.

[51] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci., USA*, 83:9373–9377, 1986.

[52] M. S. Waterman. Secondary structure of single - stranded nucleic acids. *Studies on foundations and combinatorics, Advances in mathematics supplementary studies, Academic Press N.Y.*, 1:167 – 212, 1978.

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

[54] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chemie*, 125:167–188, 1994.

[55] M. Tacker, P. F. Stadler, E. G. Bornberg-Bauer, I. L. Hofacker, and P. Schuster. Algorithm independent properties of RNA secondary structure prediction. *Eur. Biophys. J.*, 25:115–130, 1996.

[56] I. L. Hofacker, P. Schuster, and P. F. Stadler. Combinatorics of RNA secondary structures. *Discr. Appl. Math.*, 1996. submitted, SFI preprint 94-04-026.

[57] P. F. Stadler and C. Haslinger. RNA structures with pseudo-knots: Graph-theoretical and combinatorial properties. *Bull. Math. Biol.*, 1997. submitted, Santa Fe Institute Preprint 97-03-030.

[58] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc.Roy.Soc.Lond.B*, 255:279–284, 1994.

[59] P. Schuster. How to search for RNA structures. Theoretical concepts in evolutionary biotechnology. *J. Biotechnology*, 41:239–257, 1995.

[60] W. Grüner, R. Giegerich, D. Strothmann, C. M. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks. *Monath. Chem.*, 127:355–374, 1996.

[61] W. Grüner, R. Giegerich, D. Strothmann, C. M. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structures of neutral networks and shape space covering. *Monath. Chem.*, 127:375–389, 1996.

[62] C. Chothia. Proteins. one thousand families for the molecular biologist. *Nature*, 357:543–544, 1992.

[63] L. Holm and C. Sander. Dali/FSSP classification of three-dimensional protein folds. *Nucl. Acids Res.*, 25:231–234, 1997.

[64] A. G. Murzin. New protein folds. *Curr. Opin. Struct. Biol.*, 4:441–449, 1994.

[65] A. G. Murzin. Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.*, 6:386–394, 1996.

[66] I. L. Hofacker, M. A. Huynen, P. F. Stadler, and P. E. Stolorz. Knowledge discovery in rna sequence families of HIV using scalable computers. In E. Simoudis, J. Han, and U. Fayyad, editors, *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR*, pages 20–25, Menlo Park, CA, 1996. AAAI Press.

[67] S. Rauscher, C. Flamm, C. Mandl, F. X. Heinz, and P. F. Stadler. Secondary structure of the 3'-non-coding region of flavivirus genomes: Comparative analysis of base pairing probabilities. *RNA*, 3:779–791, 1997.

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

[69] W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster. Statistics of rna secondary structures. *Biochemistry*, 33:1389–1404, 1993.

[70] A. Babajide, I. L. Hofacker, M. J. Sippl, and P. F. Stadler. Neutral networks in protein space: A computational study based on knowledge-based potentials of mean force. *Folding & Design*, 2:261–269, 1997.

[71] M. J. Sippl. Calculation of conformational ensembles from potentials of mean force — an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, 213:859–883, 1990.

[72] M. J. Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17:355–362, 1993. URL: `http://lore.came.sbg.ac.at/Extern/software/Prosa/prosa.html`.

[73] M. J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding. an approach to the computational determination of protein structures. *J. Computer-Aided Molec. Design*, 7:473–501, 1993.

[74] C. M. Reidys, P. F. Stadler, and P. Schuster. Generic properties of combinatory maps: Neural networks of RNA secondary structures. *Bull. Math. Biol.*, 59:339–397, 1997.

[75] C. M. Reidys. Random induced subgraphs of generalized n-cubes. *Adv. Appl. Math.*, 1997. in press.

[76] J. Maynard-Smith. Natural selection and the concept of a protein space. *Nature*, 225:563–564, 1970.

[77] M. A. Huynen, P. F. Stadler, and W. Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. (USA)*, 93:397–401, 1996.

[78] M. A. Huynen. Exploring phenotype space through neutral evolution. *J. Mol. Evol.*, 43:165–169, 1996.

[79] M. A. Martinez, V. Pezo, P. Marlière, and S. Wain-Hobson. Exploring the functional robustness of an enzyme by *in vitro* evolution. *EMBO J.*, 15:1203–1210, 1996.

[80] W. Hordijk. Correlation analysis of the synchronizing-ca landscape. *Physica D*, 107:255–264, 1997.

[81] M. Kimura. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK, 1983.

[82] W. Ebeling, A. Engel, B. Esser, and R. Feistel. Diffusion and reaction in random media and models of evolution processes. *J. Stat. Phys.*, 37:369–384, 1984.

[83] R. Feistel and W. Ebeling. Models of Darwinian processes and evolutionary principles. *Biosystems*, 15:291–299, 1982.

[84] J. Hofbauer and K. Sigmund. *Dynamical Systems and the Theory of Evolution.* Cambridge University Press, Cambridge U.K., 1988.

[85] L. S. Tsimring, H. Levine, and D. A. Kessler. RNA virus evolution via a fitness-space model. *Phys. Rev. Letters*, 76:4440–4443, 1996.

[86] B. Derrida and L. Peliti. Evolution in a flat fitness landscape. *Bull.Math.Biol.*, 53, 1991.

[87] P. Schuster. Landscapes and molecular evolution. *Physica D*, 107:351–365, 1997.

[88] P. Schuster. Genotypes with phenotypes: Adventures in an RNA toy world. *Biophys. Chem.*, 1997. in press, Santa Fe Institute preprint 97-04-036.

[89] C. V. Forst, C. M. Reidys, and J. Weber. Evolutionary dynamics and optimization: Neutral Networks as model-landscape for RNA secondary-structure folding-landscapes. In F. Morán, A. Moreno, J. Merelo, and P. Chacón, editors, *Advances in Artificial Life*, volume 929 of *Lecture Notes in Artificial Intelligence*, pages 128–147, Berlin, Heidelberg, New York, 1995. ECAL '95, Springer.