J. Complex Systems (1998) **1**, 39–66

Amplitude Spectra of Fitness Landscapes

Wim Hordijk
Santa Fe Institute
1399 Hyde Park Road, Santa Fe, NM 87501, USA
wim@santafe.edu

Peter F. Stadler †
Santa Fe Institute
Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
studla@tbi.univie.ac.at

(Received 23 February 1998)

ABSTRACT. Fitness landscapes can be decomposed into elementary landscapes using a Fourier transform that is determined by the structure of the underlying configuration space. The amplitude spectrum obtained from the Fourier transform contains information about the ruggedness of the landscape. It can be used for classification and comparison purposes. We consider here three very different types of landscapes using both mutation and recombination to define the topological structure of the configuration spaces. A reliable procedure for estimating the amplitude spectra is presented. The method is based on certain correlation functions that are easily obtained from empirical studies of the landscapes.
KEYWORDS: Fitness landscapes, amplitude spectrum, Fourier transform, evolutionary processes, mutation, recombination, discrete Laplace operator.

---

† Address for correspondence

1. Introduction

Evolutionary change is caused by the spontaneously generated genetic variation and its subsequent fixation by drift and/or selection. Algebraic methods and approaches have gained increasing prominence in evolutionary theory in recent years, see e.g. (Lyubich, 1992). This development has been stimulated to some extent by the application of evolutionary models to design so-called evolutionary algorithms—for instance the genetic algorithms (GAs), evolution strategies (ESs), and genetic programming (GP)—and by the developing theory of complex adaptive systems.

In general, genetic variation is generated independently from the natural selection acting on it. The structure of an evolutionary model can thus be written in the form $x' = S(\{x, w\}) + T(\{x, t\})$, where x describes the frequency distribution over types (genotypes, gametes, etc.) and $S(\{x, w\})$ is a term describing the selection forces acting on x. The parameters w determine the *fitness* of a genotype. The second term $T(\{x, t\})$ describes the transmission processes by determining the probability of transforming one type into another by mutation and/or recombination (Altenberg and Feldman, 1987). Hence, evolution models can be seen as dynamical systems which "live" on an algebraic structure determined by the genetic processes like mutation and recombination (Lyubich, 1992).

Considering the fitness parameters as an explicit function of the genotypes, Sewall Wright (1932) introduced the notion of a *fitness landscape*. Implicit in this idea is a sense of "nearness" between different genotypes that is determined by the transmission term $T(\{x, t\})$. In many applications like GAs and ESs there is an unambiguous notion of "neighborhood" which is determined by the genetic operators that are used in the algorithm. This immediately imposes a topological structure on the set of configurations.

The idea of using a simpler landscape model to approximate a more realistic (biological, physical, or combinatorial) landscape is not new. Kauffman and Weinberger, for instance, used the spin-glass like NK landscapes as a model for the *affinity landscapes* arising in some models for the maturation of immune response (Kauffman, 1989). The NK model contains a parameter k that allows one to tune the "ruggedness" of the resulting landscapes. Kauffman and Weinberger adjusted this parameter such that the lengths of adaptive walks agreed reasonably well with experimental data. Furthermore, Fontana and co-workers obtained estimates for k for RNA free energy landscapes by comparing correlation data (Fontana *et al.*, 1993). By only focusing on fitting a single parameter, however, these approaches do not allow conclusions on how well the experimental landscape is actually approximated, or which characteristic(s) of a landscape should be used for fitting the free parameter(s) in the model.

Recently an algebraic framework was proposed for the study of landscapes leading to an unambiguous decomposition of a landscape that is reminiscent of a Fourier series (Stadler, 1996). In this contribution we apply this formalism to three very different landscapes taken from different fields, namely the spin-glass like NK model, the performance landscape of the synchronizing CA problem, and the free energy landscapes of RNA folding. We shall see that a single parameter is insufficient to capture the differences between these models.

This contribution is organized as follows: Section 2 introduces the concepts and mathematical notations of fitness landscapes. In section 3 we review the algebraic theory of fitness landscapes, introduce the amplitude spectrum as an important characteristic of a landscape and explain a workable procedure for its computation. This procedure is discussed in detail using the NK model as an example in section 4. In sections 5 and 6 we analyze the synchronizing-CA landscape and a particular instance of an RNA landscape, respectively. Finally, section 7 summarizes the main conclusions and provides some discussion.

2. Fitness Landscapes

2.1. Definition of Fitness Landscapes

The notion of a *landscape* plays an important role in the theory of evolution and in optimization theory ever since this concept has been introduced by Sewall Wright (1932). A landscape is composed of three ingredients. First, there is a *representation set* V, which is countable and usually finite. We shall consider here only sets of words from some finite size alphabet, although this restriction is not necessary in general. Each *configuration* in V represents, for instance, a possible genotype in a biological system, or a candidate solution of a combinatorial optimization problem that is modeled by the landscape. As an example, V could be the set of bit strings of length n, i.e., $V = \{0,1\}^n$, or the possible RNA sequences of length n, $V = \{\mathbf{G},\mathbf{C},\mathbf{A},\mathbf{U}\}^n$.

Secondly, there is a geometrical or topological structure $\mathcal{X}$ on V that determines which configurations are "close" to each other. The representation set V together with the structure $\mathcal{X}$ is often referred to as the *configuration space*. $\mathcal{X}$ is imposed by a search process on V. It can be a neighborhood-relation implying a graph, it can be a metric, or a more general topology, or another algebraic or combinatorial construction. The two most prominent examples are mutation and crossover. The configuration space becomes a simple graph under mutations, with two configurations being connected by an edge if they can be inter-converted by a single mutation. Recombination, on the other hand, gives rise to a much more involved structure on V, which has been termed a P-structure in (Stadler and Wagner, 1998).

Finally, there is a fitness (or cost) function $f : V \to \mathbb{R}$ that assigns a real value to each configuration in V. For example, the function f could be the objective function of some optimization problem. Note that our discussion here always refers to a particular function f, i.e., to a particular instance of a problem. Ensembles of landscapes are discussed in detail in (Binder and Young, 1986; Mézard *et al.*, 1987), and the cost functions of combinatorial optimization problems (Garey and Johnson, 1979) have the same abstract structure. The basis of our approach is an algebraic encoding of the configuration space that will allow us to speak about landscapes in terms of eigenvalues and eigenvectors.

2.2. Mutation: Landscapes on Graphs

Consider mutation based search. In this case there is a rule that allows us to construct "neighboring" configurations from a given one. Hence the topological structure $\mathcal{X}$ of V is determined by a relation E determining which pairs of configurations (x, y) are neighbors. Oftentimes E is symmetric, i.e., x can be obtained from y by a single mutation step if and only there is single mutation converting x back into y. (If one allows insertions and deletions, the relation is not necessarily symmetric). A symmetric relation E may be regarded as the edge set of a graph Γ with vertex set V. The configuration space of a mutation driven procedure may thus be regarded as a simple undirected graph.

The most straightforward algebraic encoding of a graph $\Gamma = (V, E)$ is its *adjacency matrix* $\mathbf{A}$, which has the entries

$$\mathbf{A}_{xy} = \begin{cases} 1 & \text{if } \{x, y\} \in E, \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Hence the entry $\mathbf{A}_{xy}$ equals 1 if and only if there is an edge in E connecting the vertices x and y. The number of neighbors of a vertex x, i.e., the number of edges incident with x, is the *degree* of x. We will repeatedly encounter the *degree matrix* $\mathbf{D}$, that is, the diagonal of the vertex degrees: $\mathbf{D}_{xx} = \sum_y \mathbf{A}_{xy}$. In this contribution we will only encounter *regular* graphs, i.e., all vertices $x \in V$ have the same degree $D = \mathbf{D}_{xx}$, since we restrict ourselves to point-mutation.

From the mathematical point of view it is more natural to characterize a graph by its *Laplacian* matrix

$$-\Delta = \mathbf{D} - \mathbf{A}. \tag{2.2}$$

For regular graphs this definition simplifies to $\Delta = \mathbf{A} - D\mathbf{I}$. The graph Laplacian $-\Delta$ is a natural generalization of the more familiar Laplacian differential operator in continuous spaces. For a detailed discussion see (Mohar, 1991; Stadler, 1996). Chung (1997) uses a somewhat different version of the graph Laplacian, $\mathbf{L} = -\mathbf{D}^{-1/2}\Delta\mathbf{D}^{-1/2}$. For a regular graph we have simply $\mathbf{L} = -(1/D)\Delta = \mathbf{I} - (1/D)\mathbf{A}$, i.e., the spectral properties of $\mathbf{L}$ and $-\Delta$ are the same apart from the trivial normalization factor D.

2.3. Recombination: Landscapes on P-Structures

The search in genetic algorithms works in a different manner than mutation-based search: two configurations from a population are chosen and a so-called crossover (or recombination) operator creates "offspring" from the two "parents". A convenient algebraic representation makes use of so-called *P-structures*, that is, pairs $(V, \mathcal{R})$ where $\mathcal{R} : V \times V \to \mathcal{P}(V)$ maps a pair of parents into the set of possible offspring (Stadler and Wagner, 1998; Wagner and Stadler, 1998). P-structures may be regarded as generalizations of hypergraphs (Berge, 1989). The latter were used in (Gitchoff and Wagner, 1996) to represent recombination spaces.

A P-structure can be identified by its (unweighted) *incidence matrix* $\mathbf{H}$, defined component-wise as

$$\mathbf{H}_{x,(y,z)} = \begin{cases} 1 & \text{if } x \in \mathcal{R}(y,z), \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

The number of different possible offspring of the parents y and z is

$$\eta(y,z) = |\mathcal{R}(y,z)| = \sum_{x \in V} \mathbf{H}_{x,(y,z)}. \tag{2.4}$$

From the incidence matrix we define the $V \times V$ square matrix

$$\mathbf{S}_{xy} = \frac{1}{V} \sum_{z \in V} \mathbf{H}_{x,(y,z)} \eta(y,z)^{-1}, \tag{2.5}$$

and finally the *P-structure Laplacian*

$$-\Delta = \mathbf{I} - \mathbf{S}. \tag{2.6}$$

We remark that the normalization of the P-structure Laplacian is the same as for Chung's version of the graph Laplacian $\mathbf{L}$ (Chung, 1997).

The mathematical details of the algebraic theory of fitness landscapes are discussed in (Stadler, 1995; Stadler, 1996; Stadler and Happel, 1998) for landscapes on graphs and in (Stadler and Wagner, 1998; Wagner and Stadler, 1998) for recombination landscapes. In section 3 we present some of the main ideas and results with as little technical overhead as possible.

2.4. Ruggedness

The notion of *ruggedness* is closely related to the difficulty of optimizing (or adapting) on a given landscape. It depends obviously on both the fitness function $f : V \to \mathbb{R}$, and on the geometry of the search space, which is induced by the search process. Two approaches for measuring this ruggedness have been considered for the case of landscapes on graphs. Palmer (1991) uses the number of local minima, that is, of configurations $\hat{x}$ with the property that $f(\hat{x}) \le f(y)$ holds for all y in the neighborhood of $\hat{x}$. Sorkin (1988), Eigen *et al.* (1989), and Weinberger (1990) use correlation measures. While the notion of locality may be ill-defined for many topological structures $\mathcal{X}$, for instance in the case of P-structures, it is still possible to define correlation measures for the landscape (Stadler and Wagner, 1998).

3. Approximation Theory for Fitness Landscapes

3.1. Fourier Analysis of Landscapes

A series expansion of a function in terms of a complete and orthonormal system of eigenfunctions $\mathbf{\Phi} = \{\varphi_k\}$ is commonly called a *Fourier expansion.* Following Weinberger (1991) the same terminology is adopted for fitness landscapes. Thus we call

$$f(x) = \sum_{k=0}^{|V|-1} a_k \varphi_k(x), \qquad x \in V \tag{3.1}$$

a *Fourier expansion* of the landscape f, and the parameters a_k are referred to as the *Fourier coefficients.* The φ_k are eigenfunctions of the Laplace operator $-\Delta$ of the landscape.

As an example consider the Boolean Hypercube $\mathcal{Q}_2^n$. This graph is regular with vertex degree $D = n$. Its vertices are the n-dimensional vectors with entries $+1$ or -1. Two such vectors are neighbors of each other if they differ in the sign of a single component. The Laplacian of $\mathcal{Q}_2^n$ has the $n+1$ eigenvalues $\lambda_p = 2p$, $p = 0, \ldots, n$. The multiplicity of λ_p is $\binom{n}{p}$. The eigenspace belonging to λ_p is spanned by the functions

$$\varphi_{i_1 i_2 \ldots i_p}(x) = 2^{-n/2} x_{i_1} x_{i_2} \ldots x_{i_p} \tag{3.2}$$

for all combinations of the p indices satisfying $1 \le i_1 < i_2 < \ldots < i_p \le n$. The collection of all these eigenfunctions, together with the constant functions $\varphi_0(x) = 2^{-n/2}$ belonging to λ_0, forms the Fourier basis $\mathbf{\Phi}$. It is easy to verify that these vectors φ_i are normalized and pairwise orthogonal. They are closely related to the probably more familiar Walsh functions (Goldberg, 1989).

As a second example consider 1-point crossover on binary strings of length n. Again we use the alphabet $\{+1, -1\}$ for convenience. As shown by Wagner and Stadler (1998) the Walsh functions, equ.(3.2), are eigenfunctions of the corresponding P-structure Laplacian with eigenvalues $\lambda_0^{[1]} = 0$ and

$$\lambda^{[1]}_{i_1,i_2,\ldots,i_p} = 1 - \frac{1}{2}\frac{n - (i_p - i_1) - 1}{n-1}. \tag{3.3}$$

The eigenvalues depend on the separation $\ell = i_p - i_1 + 1$ of the interacting sequence positions instead of the number of interacting positions in this case. The fact that the Laplacian of the Hamming graph $\mathcal{Q}^n_\alpha$ and the P-structure Laplacian associated with 1-point crossover have a common eigenbasis (and hence commute) hints at a close relationship of these objects that is explored in some detail in (Culberson, 1995; Gitchoff and Wagner, 1996; Stadler and Wagner, 1998).

Considering $\mathbf{\Phi}$ as a matrix with entries $\mathbf{\Phi}_{xk} = \varphi_k(x)$ we may use matrix notation and write a Fourier series as $f = \mathbf{\Phi}a$. The Fourier transform takes the form $a = \mathbf{\Phi}^* f$ and produces a set (or a vector) containing $|V|$ coefficients. Thus, not much is gained at first glance.

3.2. Elementary Landscapes and the Amplitude Spectrum

In 1-d Fourier analysis there are two eigenfunctions to each eigenvalue $\lambda_k > 0$ of $-\Delta$, namely $\sin(kx)$ and $\cos(kx)$. In many cases it is sufficient to determine the amplitude of the *k-th mode*, $|a_k|^2 + |a'_k|^2$ and to disregard the phase information contained in the two coefficients a_k and a'_k. This amounts to determining the relative importance of the different eigenspaces of $-\Delta$, while the structure of f within the eigenspaces is of little interest. Our analysis of discrete landscapes takes the same approach. The difference is, however, that the number and multiplicity of the eigenvalues of the Laplacian depends on the structure of the graph Γ. Highly symmetric configuration spaces, such as the Boolean Hypercube $\mathcal{Q}^n_2$, have only a relatively small number of distinct eigenvalues with (on average) very high multiplicities.

A constant function, $f(x) = c$, is an eigenvector of $-\Delta$ with eigenvalue $\lambda_0 = 0$. These landscapes are usually called *flat*. In the following we shall assume that f is not a flat landscape.

A landscape is *elementary* if it is of the form $f(x) = c + \varphi(x)$, where c is a constant and φ is an eigenvector of the Laplacian $-\Delta$ belonging to an eigenvalue $\lambda > 0$. It turns out that c is the average of f over all $x \in V$. It is clear that we can write an arbitrary landscape in the form

$$f(x) = c + \sum_p \beta_p \tilde{\varphi}_p(x). \tag{3.4}$$

Here $\tilde{\varphi}_p$ is a normalized eigenvector of $-\Delta$ belonging to the eigenvalue $\lambda_p > 0$. The index p runs only over the distinct eigenvalues of $-\Delta$. Of course we have $c = a_0$ and in general

$$\beta_p \tilde{\varphi}_p(x) = \sum_{k:\, \Delta\varphi_k = \lambda_p \varphi_k} a_k \varphi_k(x). \tag{3.5}$$

Multiplying this equation by its complex-conjugate and summing over all $x \in V$ immediately yields

$$|\beta_p|^2 = \sum_{k:\, \Delta\varphi_k = \lambda_p \varphi_k} |a_k|^2. \tag{3.6}$$

Equ.(3.4) shows that any landscape is a superposition of elementary landscapes.

It seems natural to consider the coefficients $|\beta_p|^2$ as a kind of amplitude spectrum. Using the following identity, which is straightforward to check,

$$\sum_p |\beta_p|^2 = \sum_{k \neq 0} |a_k|^2 = \sigma^2 = \sum_{x \in V} \big(f(x) - c\big)^2, \tag{3.7}$$

we define the (normalized) amplitudes as $B_p = |\beta_p|^2/\sigma^2$. We will call the vector $\{B_p\}$, $p = 1, \ldots, p_{\max}$, the *amplitude spectrum* of the landscape f. For later reference we remark: $B_p \geq 0$ for all $p = 1, \ldots, p_{\max}$, and $\sum_p B_p = 1$.

Elementary landscapes form an important class because the landscapes obtained from the most intensively studied combinatorial optimization problems with natural choices of neighborhood relations on V are elementary. Examples include the symmetric traveling salesman problem (Lawler *et al.*, 1985) with either transpositions or inversions, the graph bipartitioning problem with exchange of two vertices between the two partitions, and the graph coloring problem with mutation of the color of a single vertex, see (Codenotti and Margara, 1992; Grover, 1992). Elementary landscapes have special geometric properties. For instance, all local maxima have fitness values above the average fitness c. For a detailed discussion see (Stadler, 1996).

It is interesting to compare the amplitude spectra B_p^m for mutation based search with the spectra $B_p^{[1]}$ of the same landscape under 1-point crossover. As an immediate consequence of the explicit expressions for the amplitudes, equ.(3.6), and the fact that the Walsh functions form a Fourier basis in both cases, we observe that $B_1^m = B_1^{[1]}$. In other words, the linear component of a fitness landscape has the same influence under mutations and 1-point crossover.

Furthermore, $B_2^m \geq B_2^{[1]}$ since all contributions to $B_2^{[1]}$ have exactly two interacting bits and thus also contribute to B_2^m, while, for instance, the basis function $\varphi_{13}(x)$ and hence the correspond Fourier coefficient a_{13} contributes to B_2^m but not to $B_2^{[1]}$. Note that B_1 measures the additive fitness contributions. Indeed, a landscape with $B_1 = 1$ is a purely additive (single-peak) landscape. The higher-order contributions are responsible for the ruggedness of a landscapes. Hence one might use $1 - B_1$ as a measure of the amount of *epistasis* in the fitness function.

3.3. Correlation Functions and Ruggedness

A seemingly very different approach towards characterizing complex landscapes is based on their correlation structure. Intuitively, the correlations between fitness values of nearby points determine the *ruggedness* of a landscape. While different authors use different aspects of a landscape to define ruggedness — Sorkin (1988), Eigen *et al.* (1989), and Weinberger (1990) introduce slightly different correlation measures, Palmer (1991) proposes the statistics of local optima, Kauffman and Levin (1987) use adapative walks — all these notions are closely related, at least in generic cases (García-Pelayo and Stadler, 1997).

The simplest correlation measure (from a mathematical point of view) is based on a simple (unbiased) random walk on V. Such a walk has the transition matrix $\mathbf{AD}^{-1}$, i.e., at each vertex one of the $\mathbf{D}_{xx}$ (Hamming) neighbors is chose with uniform probability. Recording the fitness of each point encountered along the walk gives rise to a "time series" of fitness values $\{f_0, f_1, \dots f_T\}$. In (Hordijk, 1996) a similar random walk procedure for crossover was suggested. Starting with an arbitrary pair of parents y and z, crossover produces a set of offspring, namely $\mathcal{R}(y,z)$ in our P-structure formalism. From those we choose one with equal probability $1/\eta(y,z)$. In the next generation this offspring y' is mated with a random element of V. Clearly, $\mathbf{S}$ is the transition matrix of the resulting random walk (Stadler and Wagner, 1998; Wagner and Stadler, 1998).

Up to second order, the properties of this "time series" of fitness values are determined by

$$E[f_t] = \frac{1}{|V|}\sum_{x\in V} f(x) \qquad \text{and} \qquad \mathrm{Var}[f_t] = \frac{1}{|V|}\sum_{x\in V}\left(f(x) - E[f_t]\right)^2, \tag{3.8}$$

and the autocorrelation function

$$r(s) = \frac{E[f_t f_{t+s}] - E[f_t]E[f_{t+s}]}{\mathrm{Var}[f_t]}. \tag{3.9}$$

A key result of the algebraic theory of fitness landscapes (Stadler, 1996) states that, for any regular graph Γ and any non-constant fitness functions f,

$$r(s) = \sum_{p \neq 0} B_p (1 - \lambda_p / D)^s. \tag{3.10}$$

The same result holds for non-flat landscapes on the P-structures derived from string-recombination. Due to the somewhat different normalization of the P-structure Laplacian, equ.(2.6), we have $D = 1$ in the case of crossover.

The autocorrelation function $r(s)$ is therefore uniquely determined by the structure of the graph or P-structure that underlies the search process (through the eigenvalues λ_p of its Laplacian) and amplitude spectrum of the fitness function. In particular, if f is an elementary landscape then its autocorrelation function $r(s)$ is a single exponential. Detailed numerical studies of $r(s)$ for the landscapes of the correlation structure of a variety of combinatorial optimization problems with both elementary and non-elementary landscapes are reported e.g. by Stadler and Schnabl (1992), Stadler and Happel (1992), Stadler (1992), and Angel and Zissimopoulos (1997, 1998a, 1998b).

3.4. Estimating Amplitudes

The direct computation of the amplitude spectrum using the Fourier transform of the fitness function is restricted to small problems since it requires the knowledge of the fitness value of every point in the landscape. By combinatorial nature of configuration spaces, $|V|$ usually increases exponentially with the system size (e.g., the number of cities in a TSP, the length of a DNA molecule, or number of interacting spins), making this approach infeasible for problem sizes of practical importance. In many cases the computational efforts for the Fourier transform is prohibitive even when the fitness function can be evaluated exhaustively. In some cases, such as the Boolean hypercube, this limitation is partially alleviated by the availability of Fast Fourier Transform techniques (Welch, 1968; Cairns, 1971; Maslen and Rockmore, 1996; Rockmore, 1995). We may, however, estimate the amplitude spectrum from randomly sampled points of the landscape.

The autocorrelation function $r(s)$ of a fitness landscape f can be estimated easily by sampling data along a random walk:

$$\hat{r}(s) = \frac{\sum_{t=1}^{T-s} (f_t - \overline{f})(f_{t+s} - \overline{f})}{\sum_{t=1}^{T} (f_t - \overline{f})^2}, \qquad \text{where} \qquad \overline{f} = \frac{1}{T} \sum_{t=1}^{T} f_t. \tag{3.11}$$

Provided the eigenvalues λ_p of the underlying graph Γ are known, we can use the estimated autocorrelations $\hat{r}(s)$ to estimate the values of the amplitudes B_p using equ.(3.10). A first attempt in this direction was using a different type of correlation function (Happel and Stadler, 1996).

4. An Illustrative Example: NK Landscapes

Three very different kinds of landscapes are considered in this paper. The first one, an instance of Kauffman's NK model (Kauffman and Levin, 1987), is spin-glass like and similar to the simply-stated, well-studied combinatorial optimization problems such as the TSP or graph bipartitioning. The second example is the synchronizing-CA landscape (Das *et al.*, 1995; Hordijk, 1997) representing the type of landscapes often encountered in the complex optimization problems of technical importance. The third example, RNA folding landscapes, is taken from biophysics. As an illustration of our approach we carry out the analysis of a (very) small example in full detail in this section.

Kauffman's NK model defines a family of fitness landscapes that can be tuned from smooth via rugged to completely random (Kauffman, 1989; Kauffman, 1993). The configurations in the landscape are bit strings of length n, i.e., $V = \{0,1\}^n$. Each individual bit b_i in a string $\mathbf{b}$ has a fitness contribution f_i which not only depends on its own value (0 or 1), but also on the values of k other bits in the string. Each bit basically has a lookup table with 2^{k+1} entries, one for each possible configuration of the bit itself and its k "neighbors". Each entry in this lookup table is assigned a random value between 0 and 1. These assignments are done independently for each bit in the string. Once these assignments are made, however, they remain fixed, giving rise to a particular instance of an NK landscape.

The fitness of a bit string is calculated by first determining the fitness contribution f_i of each bit b_i. The value of f_i, which depends on the bit b_i and its k neighbors, is contained in the lookup table. The fitness of the entire string is then simply the average of the fitness contributions of all the bits in the string:

$$f(\mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} f_i. \tag{4.1}$$

The ruggedness of NK landscapes can be tuned from a completely smooth, single-peaked landscape ($k = 0$) to a completely random, multi-peaked one ($k = n-1$).

There are several ways in which the k neighbors for each bit can be chosen. The most common choice is *nearest neighbor interactions*, i.e., one takes the $k/2$ bits directly to the left and to the right of the bit i itself (assuming that k is even). For a more detailed description of the model, its variants, and its applications we refer to the book (Kauffman, 1993).

Let us now turn to a simple numerical example. Consider an instance of the NK model with $n = 5$ and $k = 2$. There are only $|V| = 2^n = 32$ points in this landscape. The fitness values for our particular example are listed in table 1.

The Fourier basis $\mathbf{\Phi}$ for both the Boolean hypercube of the mutation landscape

Table 1. The fitness values of the 32 points in the particular instance of an NK landscape as used here.

x	$f(x)$	x	$f(x)$	x	$f(x)$	x	$f(x)$
00000	0.703	01000	0.663	10000	0.432	11000	0.402
00001	0.497	01001	0.516	10001	0.524	11001	0.517
00010	0.640	01010	0.667	10010	0.409	11010	0.446
00011	0.510	01011	0.596	10011	0.300	11011	0.360
00100	0.757	01100	0.690	10100	0.596	11100	0.416
00101	0.478	01101	0.470	10101	0.616	11101	0.459
00110	0.649	01110	0.664	10110	0.528	11110	0.431
00111	0.680	01111	0.755	10111	0.581	11111	0.506

and the P-structure of the crossover landscape consists of the eigenfunctions defined in equ.(3.2). For each eigenfunction $\varphi_{i_1 i_2 \ldots i_p}$ we may represent the set of indices $i_1, i_2, \ldots, i_p$ by a bit string $\mathbf{i}$ of length n with 1s at the bit positions $i_1, i_2, \ldots, i_p$ and 0s at the other positions. A natural ordering of the 2^n eigenfunctions is obtained by interpreting the bit string $\mathbf{i}$ as the binary representation of an integer i. Note that $0 \leq i < 2^n$. It is now straightforward to verify that we may compute $\varphi_i(x)$ very efficiently using the representation

$$\varphi_i(x) = 2^{-n/2} \xi(i \wedge \overline{x}) \tag{4.2}$$

where $\wedge$ is the bitwise `AND` operator and $\overline{x}$ is the complement of the bit string x. $\xi(y)$ of a bit string y is $+1$ or -1, depending on whether the number of 1s in y is even or odd, respectively. As an example, we have,

$$\varphi_{21}(28) = \varphi_{\texttt{'10101'}}(\texttt{'11100'}) = 2^{-5/2}\xi(\texttt{'00001'}) = (-1)2^{-5/2}.$$

Using eqn.(4.2), the Fourier transform $a = \mathbf{\Phi}^* f$ produces the coefficients listed in table 2.

Once the Fourier coefficients a_i are known, we can directly compute the amplitudes B_p. Recall that for the mutation landscape the eigenfunction ρ_i, and thus the Fourier coefficient a_i, belongs to the eigenvalue λ_p if the binary representation $\mathbf{i}$ of the index i contains exactly p 1s. Similarly, for the crossover landscape the Fourier coefficient a_i belongs to the eigenvalue λ_ℓ if in the binary representation $\mathbf{i}$ of i the separation (i.e., the distance between the leftmost and rightmost 1 plus one) is equal to ℓ. The values for the number of 1s p and the separation ℓ for the indices i of the Fourier coefficients a_i are given in table 2. So, as an example, to calculate the value of B_2 we need the a_i with indices $i \in \{3, 5, 6, 9, 10, 12, 17, 18, 20, 24\}$ for the mutation landscape, and

Table 2. Fourier coefficients a_i (above) and amplitudes B_p (below) for the NK landscape of table 1. i gives the index of the Fourier coefficients a_i, $\mathbf{i}$ is the binary notation of the index i, p is the number of 1s in $\mathbf{i}$, and ℓ is the separation, i.e., the distance between the leftmost and rightmost 1s in $\mathbf{i}$ plus one. p and ℓ are used to calculate the amplitudes for the mutation and crossover case respectively.

i	$\mathbf{i}$	p	ℓ	a_i	i	$\mathbf{i}$	p	ℓ	a_i
0	00000	0	0	3.08617	16	10000	1	1	-0.42639
1	00001	1	1	-0.12869	17	10001	2	5	0.20047
2	00010	1	1	-0.00247	18	10010	2	4	-0.13930
3	00011	2	2	0.07707	19	10011	3	5	-0.19622
4	00100	1	1	0.19339	20	10100	2	3	0.06930
5	00101	2	3	0.06293	21	10101	3	5	0.00035
6	00110	2	2	0.11278	22	10110	3	4	-0.00000
7	00111	3	3	0.16546	23	10111	4	5	-0.00035
8	01000	1	1	-0.06046	24	11000	2	2	-0.09829
9	01001	2	4	0.05798	25	11001	3	5	-0.02581
10	01010	2	3	0.10571	26	11010	3	4	0.00000
11	01011	3	4	-0.00000	27	11011	4	5	-0.00035
12	01100	2	2	-0.11420	28	11100	3	3	-0.08697
13	01101	3	4	0.00000	29	11101	4	5	-0.00035
14	01110	3	3	0.01096	30	11110	4	4	-0.00000
15	01111	4	4	0.00000	31	11111	5	5	-0.00035

	B_1	B_2	B_3	B_4	B_5
Mutation	0.54675	0.28374	0.16951	0.00000	0.00000
Crossover	0.54675	0.09445	0.12560	0.05199	0.18121

$i \in \{3, 6, 12, 24\}$ for the crossover landscape. The amplitude spectra for both the mutation and crossover landscapes are listed in table 2 (lower part).

Note that for the mutation landscape only the first three amplitudes B_1 through B_3 have a non-zero value. This agrees with the result derived by Stadler and Happel (1998) showing that only the first $k + 1$ modes contribute to the amplitude spectrum of an NK landscape with nearest neighbor interactions and mutation only. Furthermore, as mentioned in section 3, the value of B_1 is indeed equal for the mutation and crossover landscapes.

Let us now turn to estimating the amplitudes from sample data. To this end we generate two simple random walks of 10,000 steps. In the first walk, point-mutation is used as operator, i.e., at each step a randomly chosen bit in the current string is flipped. In the second walk, the random-mate random-child

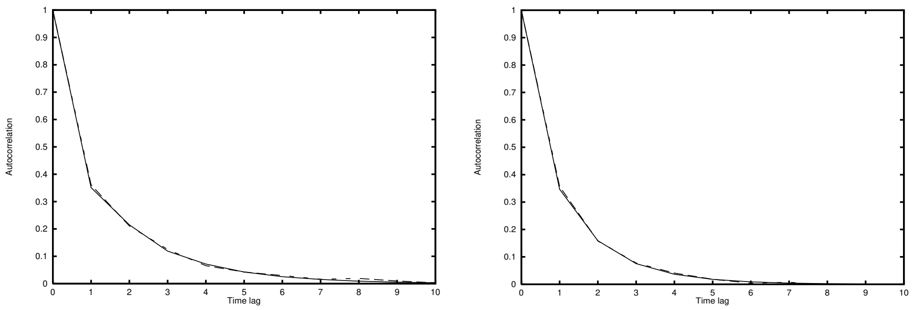

Figure 1. The autocorrelation function for mutation (l.h.s.) and crossover (r.h.s) of the NK landscape defined in table 1. Exact (solid line) and estimated (dashed line) values are compared.

one-point crossover operator is used, as introduced by Hordijk (1996). With this operator, at each step a one-point crossover is performed between the current string and a randomly chosen mate, and one of the two offspring is chosen, with equal probability, to become the new current string. Along the walk, the fitness value at each point is evaluated, and these values are used to compute the estimate $\hat{r}(s)$, equ.(3.11), of the autocorrelation function $r(s)$.

Figure 1 compares the analytical expression $r(s)$ with the estimate $\hat{r}(s)$, for both mutation (l.h.s.) and crossover (r.h.s). Not surprisingly, the agreement is excellent for such a long walk on a very small graph.

Since the eigenvalues of the Laplacian $-\Delta$ are known ($\lambda_p = 2p$ in the case of the Boolean hypercube and $\lambda_\ell = 1 - \frac{n-\ell}{2(n-1)}$ in the case of the P-structure) we may estimate the amplitudes B_p via

$$r(s) = \sum_{p \neq 0} B_p (1 - 2p/n)^s \tag{4.3}$$

for the mutation case, and via

$$r(s) = \sum_{\ell \neq 0} B_\ell \left(1 - \frac{n-\ell}{2(n-1)}\right)^s \tag{4.4}$$

for the crossover case by substituting the respective $\hat{r}(s)$ values for the $r(s)$. This leads to a system of linear equations

$$\begin{array}{rcrcrcccr} \hat{r}(0) & = & B_1 & + & B_2 & + & \dots & + & B_n \\ \hat{r}(1) & = & (1-2/n)B_1 & + & (1-4/n)B_2 & + & \dots & + & (1-2)B_n \\ \vdots & & \vdots & & & & \vdots & \vdots & \\ \hat{r}(k) & = & (1-2/n)^k B_1 & + & (1-4/n)^k B_2 & + & \dots & + & (1-2)^k B_n \end{array} \tag{4.5}$$

for the mutation case and a similar set of equations for the crossover case.

A straightforward solution of equ.(4.5) using a least squares procedure does not yield satisfactory results, however, as it frequently produces negative values for B_p, contradicting the constraint $B_p \geq 0$. We therefore resort to a steepest descent technique (Burden and Faires, 1989) that iteratively minimizes the sum of squared errors in the above set of equations. We impose the constraint that the variables that are to be estimated (in this case the B_p's) are not allowed to become negative.

The accuracy of a steepest descent algorithm depends strongly on certain parameter values. First of all, the number of equations $k+1$ in the above set of linear equations is important. It turns out that the best results are obtained here when $k+1$ is equal to the number of variables n to be estimated. If $k+1$ is smaller than n, the system is not fully determined. On the other hand, the $r(k)$ values are very small for k larger than n and hence the estimates $\hat{r}(k)$ contain mostly noise. At the same time, the coefficients $(1-2p/n)^k$ are small, so that the noise in $\hat{r}(k)$ has an amplified effect on the B_p's.

Furthermore, the number of iterations that the algorithm is run for will influence the final outcome. If the algorithm is not run long enough, the final result might not be accurate enough. If the algorithm is run too long, on the other hand, the parameters might be "over-estimated". Usually, a certain tolerance `tol` is used as a stopping criterion. If the possible improvement, i.e., the decrease in the objective value, becomes smaller than this tolerance, the algorithm stops. Here, we use `tol`$= 0.00001$.

The final result of the estimation procedure is also sensitive to the starting point that is used. This usually means that the algorithm needs to be run several times, each time with a different starting value. In most cases, however, the sum of squared errors differs significantly between different final results. It can thus be used to identify an acceptable solution among different runs that got stuck in local optima.

With this steepest descent algorithm the amplitudes B_p for the NK landscape are estimated using the first five estimated autocorrelations $\hat{r}(s)$ of figure 1. In the mutation case the results are insensitive to the starting points. The left hand side of figure 2 shows the exact amplitudes as given in table 2 (bars) and the

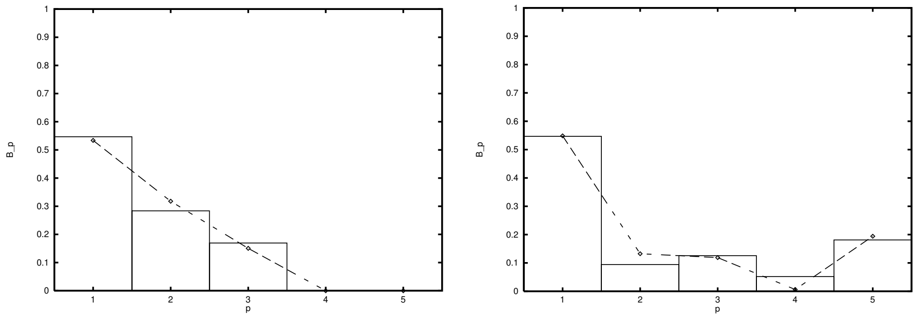

Figure 2. The amplitude spectra for mutation (l.h.s.) and crossover (r.h.s) of the NK landscape defined in table 1. Exact (bars) and estimated (dashed lines) values are compared.

estimated ones (dashed lines) for mutation. The right hand side of figure 2 shows the same data for crossover. Here, the results are sensitive to the starting points, mainly because only the first three autocorrelation values are significantly above the noise level. This makes the estimation of five parameters more difficult. The figure shows the best estimation result obtained, which resulted from a starting point that was already relatively close to the exact solution, but which clearly resulted in the lowest value of the objective function (i.e., the sum of squared errors).

5. The synchronizing-CA landscape

The next landscape we will look at is derived from the *global synchronization* task for cellular automata (CAs). The global synchronization problem consists of finding CA rules that are capable of performing a certain computational task, namely the global synchronization of an arbitrary initial pattern (Das *et al.*, 1995). Consider a 1-dimensional lattice with N identical cells and an update rule ϕ defining how $s_i(t)$, the state of cell i at time t, depends on the states of the $2r+1$ cells $s_{i-r}(t-1)$,...,$s_{i+r}(t-1)$ at the previous time step, where r is the *radius* of the CA. We assume periodic boundary conditions, i.e., $s_{i+N}(t) = s_i(t)$. The update rule ϕ is a Boolean function $\phi : \eta \in \{0,1\}^{2r+1} \to \{0,1\}$. Since each of the 2^{2r+1} possible substrings η yields either a 0 or 1 in the next iteration, there are a total of $2^{2^{2r+1}}$ different rules for a fixed radius r. Using the lexicographical ordering of the 2^{2r+1} possible bit strings η we can represent each radius r CA rule by the string $\psi = (\phi(\eta_1), \phi(\eta_2), \ldots, \phi(\eta_{2^{2r+1}}))$.

The goal of the synchronizing-CA problem is to find rules ϕ such that each initial configuration (IC) $s(0)$ reaches (after at most a finite number M steps) a period-two behavior, oscillating between an all-0s configuration and an all-1s configuration, i.e., for all cells i, $s_i(T+2k)=0$ and $s_i(T+2k+1)=1$ for some $T \leq M$ and $k = 0, 1, 2, \ldots$. Most CA rules reach such a final state only for a small fraction of the possible ICs, or not at all. Therefore, in (Das *et al.*, 1995) the fitness of a CA is defined as the fraction of a number I of ICs on which the CA reaches the correct synchronizing behavior. Here, the fitness of a CA is calculated over a set of $I = 1000$ randomly generated ICs. A new set of random ICs is generated for each fitness measurement. As in (Das *et al.*, 1995), the other parameter values used here are $N = 149$, $M = 2.15N$, and $r = 3$ (giving rise to bit strings of length 128).

In (Hordijk, 1997) two subspaces in the fitness landscape that results from the synchronizing-CA problem were investigated. These subspaces are based on the particular "particle-strategy" that a CA uses to solve the synchronization task. Briefly, most of the bits in the lookup table of a particular CA are fixed, except for those bits for which the given strategy (and corresponding fitness) of the CA does not change significantly when they are mutated. The two subspaces are then defined by the remaining "free" bits in two particular CAs (a total of 14 and 11, respectively). This is a first-order approximation, in the sense that bits were mutated individually, and not in pairs or higher-order combinations. See (Hordijk, 1997) for more details about how exactly the subspaces are defined. We are using them here since they are still small enough to be amenable to an exact amplitude spectrum calculation.

In a similar way as in the previous section, the amplitude spectra of both these subspaces in the synchronizing-CA landscape are calculated exactly and estimated via the estimated autocorrelations (not shown here), both for mutation and crossover. Figure 3 shows the results for both the first subspace (of dimension 14) and the second subspace (of dimension 11). While we find that the estimates depend on the initial values of the steepest descent procedure in this case, the results are in general very similar to the exact values, and the residual square error can be used to discriminate the true solution from local optima.

In (Hordijk, 1997) the correlation functions of the CA landscapes were estimated as an AR(2) process. Translating the results back into a random walk correlation function one finds for the first subspace

$$r(s) \approx 0.88 \times 0.70^s + 0.12 \times (-0.18)^s \,,$$

i.e., a superposition of a large smooth and a relatively small (12%) but very rugged component. The value 0.70 is the basis of the first exponential term and corresponds approximately to the $p = 2$ mode, while -0.18 corresponds to an effective value of $p \approx 8.3$. Given the large number of components that

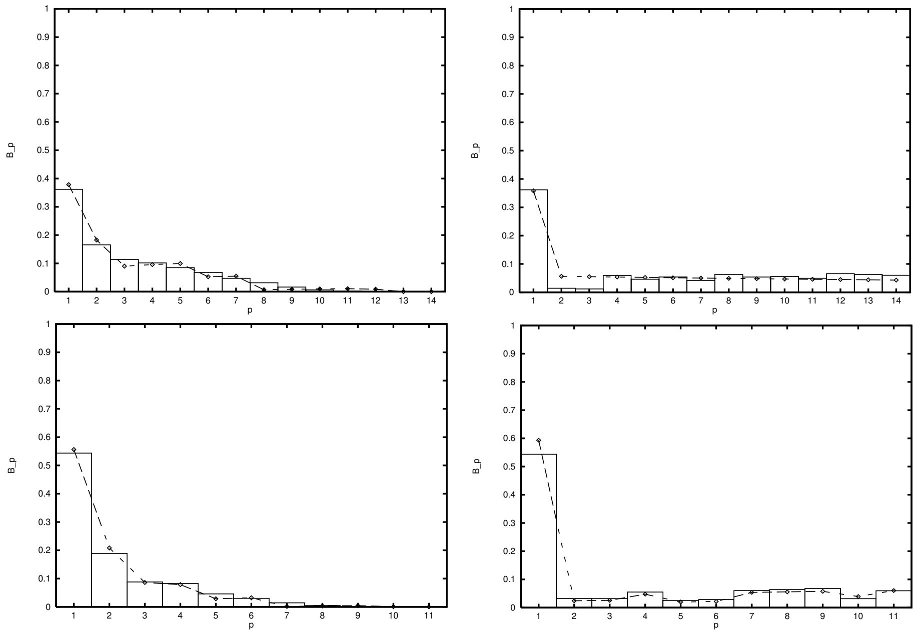

Figure 3. Amplitude spectra of the first subspace (of dimension 14; top) and the second subspace (of dimension 11; bottom) in the synchronizing-CA landscape. L.h.s.: mutation, r.h.s.: crossover. Bars indicate the exact results and the dashed lines indicate the numerical estimates.

significantly contribute to the amplitude spectrum it is not surprising that an AR(2) approximation gives only very limited information.

It is to be expected that those bits of the CA rule that have a small individual influences on the fitness will contribute predominantly additively to the fitness function. Not surprisingly, we therefore find large values of B_1 in both subspaces. On the other hand, the non-linear contributions consists of a broad distribution of high-order modes, reflecting the enormous degree of ruggedness encountered in the full landscape (Hordijk, 1997). In this sense the subspaces of the evolving CA landscapes are much more rugged than the $k = 2$ NK landscape despite the larger value of B_1.

6. RNA landscapes

Our final example consists of an RNA landscape. RNA sequences are strings over the alphabet $\{\mathbf{G},\mathbf{C},\mathbf{A},\mathbf{U}\}$ designating the nucleotides guanine, cytosine,

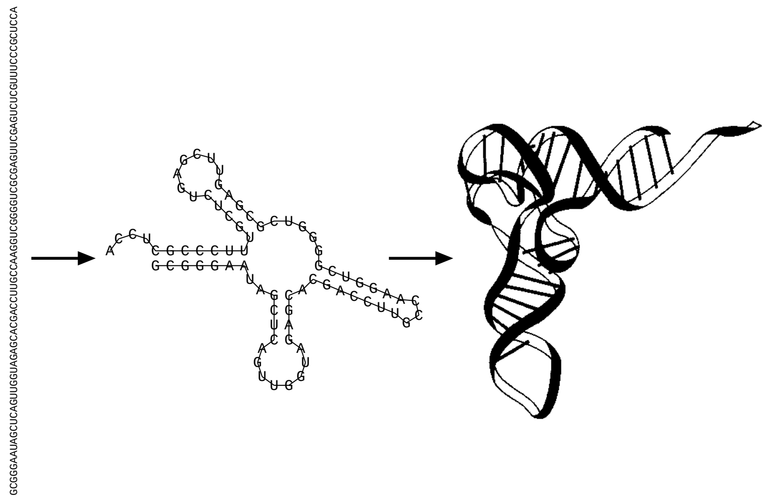

Figure 4. The folding of the one-dimensional primary structure (sequence) of an RNA molecule is decomposed conceptually into two steps: (i) Folding of the sequence into a planar secondary structure by formation of complementary Watson-Crick base pairs **GC** and **AU** as well as the weaker **GU** pairs. (ii) Formation of the 3D tertiary structure from the planar pattern. The reason for this decomposition is that the intramolecular forces stabilizing the secondary structures — base pairing and base pair stacking — are much stronger than those accounting for the particular spatial arrangement. Thus the free energy for the formation of the entire spatial structure can be estimated by the free energy for the formation of the secondary structure. While the computation of 3D structures from the pure sequence information is impossible at present, the secondary structure can be readily calculated.

adenosine, and uracil. These nucleotides form three types of stable base pairs, **GC**, **AU**, **GU** (in the order of decreasing strength). The spatial structure of an RNA molecule is approximated by its *secondary structure*, that is, the list of the base pairs minimizing the energy on the structure. The dominant role of the secondary structure is well documented in nature by the conservation of secondary structure elements in evolution (Cech, 1988). There are efficient algorithms for the prediction of the secondary structure of RNA molecules (Zuker and Sankoff, 1984; Hofacker *et al.*, 1994), that allow a detailed analysis of fitness landscapes derived from RNA structures, see e.g., (Schuster *et al.*, 1997) and the references therein. Figure 4 provides and example.

A first attempt to compute a (low resolution) amplitude spectrum of the RNA free energy landscape can be found in (Happel and Stadler, 1996). For the current study we restrict ourselves to RNA sequences consisting only of two different nucleotides, **C** and **G**. This allows a binary representation and hence a direct

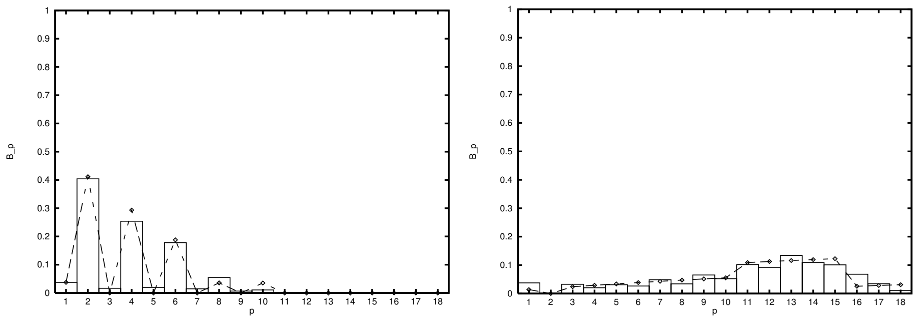

Figure 5. The exact (bars) and estimated (dashed lines) amplitude spectra for the RNA free energy landscape with $n = 18$ for mutation (l.h.s.) and crossover (r.h.s.). As in the previous examples we compare the exact values with the best found estimates.

comparison of the RNA landscape with the NK model and the synchronizing-CA landscape. The computation of the RNA secondary structures and their energies of folding was performed using the `Vienna RNA Package`, release 1.2 (Nov. 1997). This software is freely available on the internet†.

An exact computation of the amplitude spectrum was performed for a chain length of $n = 18$. The results are shown in figure 5. The limiting factor here is the Fourier transform, not the RNA folding. In fact, the secondary structures of all **GC** sequences with chain length up to $n = 30$ have been computed and analyzed by Grüner *et al.* (1996a, 1996b). For the estimation of the amplitude spectrum we used random walks with 100,000 steps. Walks with 10,000 steps give very inaccurate estimates for the autocorrelation function $r(s)$, resulting in poor estimates for the amplitudes.

The steepest descent method faces formidable problems with the RNA data. We found a strong dependence of the results on the starting points. Good estimates (with reasonably small residual sum of squared errors) can be found only when the starting point is not too far from the exact values of the amplitudes. Only the even modes play a significant role, while the amplitudes of the odd modes are close to 0. We find that starting points with equal values of the even modes and vanishing odd modes give good estimates, while starting points with substantial odd-mode contributions do not converge well and/or get stuck in local optima.

As an example, table 3 shows the results of the amplitude spectrum estimations for the RNA landscape for mutation, starting with four different starting points.

† `http://www.tbi.univie.ac.at/`

Table 3. The amplitude spectrum estimation results for the RNA landscape for mutation with four different starting points. Only the first 10 amplitudes are shown here. The starting points and the final results are shown, together with the final value of the objective function $g(x)$, i.e., the sum of squared errors. $B_p < 0.002$ for $p > 10$.

	start	final	start	final	start	final	start	final	exact
B_1	0.0	0.1034	1.0	0.1089	0.0	0.0793	0.0	0.0375	0.03748
B_2	0.0	0.2071	1.0	0.1959	0.2	0.3137	0.4	0.4114	0.40388
B_3	0.0	0.1980	1.0	0.1953	0.0	0.0833	0.0	0.0000	0.01651
B_4	0.0	0.1628	1.0	0.1649	0.2	0.2389	0.3	0.2933	0.25389
B_5	0.0	0.1252	1.0	0.1299	0.0	0.0135	0.0	0.0000	0.01969
B_6	0.0	0.0910	1.0	0.0965	0.2	0.1544	0.2	0.1877	0.17830
B_7	0.0	0.0615	1.0	0.0654	0.0	0.0000	0.0	0.0000	0.01474
B_8	0.0	0.0362	1.0	0.0358	0.2	0.0869	0.05	0.0357	0.05447
B_9	0.0	0.0145	1.0	0.0065	0.0	0.0000	0.0	0.0000	0.00675
B_{10}	0.0	0.0000	1.0	0.0000	0.2	0.0342	0.05	0.0352	0.01077
$g(x)$	0.0004056		0.0005259		0.0003656		0.0000529		

The table clearly shows that in this case different starting points can give rise to completely different final values. However, once a good starting point is found, the final result is clearly distinguished by the much lower value of the objective function (in this case almost one order of magnitude smaller than the other values).

The free energy landscapes of RNA secondary structures have been studied in great detail, see for instance (Fontana *et al.*, 1991; Fontana *et al.*, 1993; Bonhoeffer *et al.*, 1993; Tacker *et al.*, 1996). Among other properties, the *correlation length*

$$\ell = \sum_{s=0}^{\infty} r(s) = D \sum_{p>0} \frac{B_p}{\lambda_p} = \frac{n}{2} \sum_{p>0} \frac{B_p}{p} \tag{6.1}$$

is considered in these studies. Thus $\overline{p} = n/2\ell$ may be interpreted as the average amplitude (in fact, this quantity is the amplitude-weighted harmonic mean of p).

Using a two-mode model of the form $f(x) = B_p\varphi_p(x) + (1 - B_p)\varphi_q(x)$, where p and q are adjusted to optimally fit the data, an estimated value of $\overline{p} \approx 6$ was obtained in (Happel and Stadler, 1996). The data in figure 6 clearly show that such a simplified model is not adequate: at least the first six even modes play a major role for $n = 100$. Not surprisingly, the detailed data reported below give

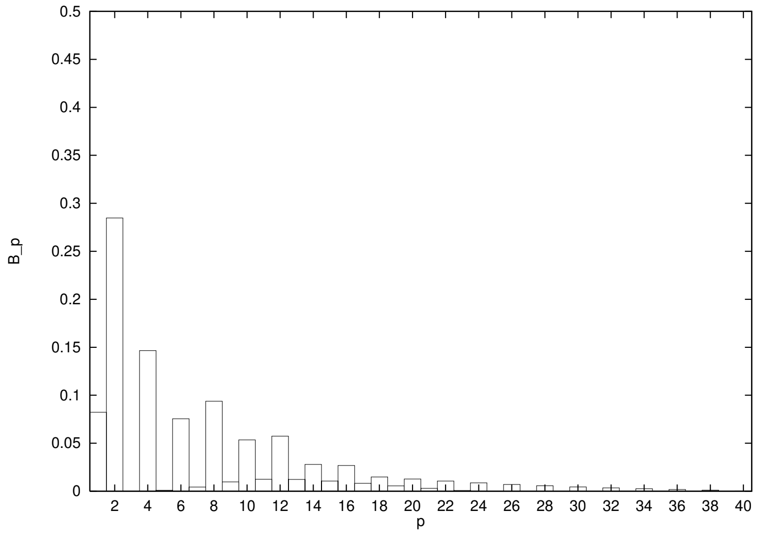

Figure 6. The estimated amplitude spectrum for a GC landscape with $n = 100$ for mutation.

a substantially different value of $\overline{p}$, see table 4. To this end we have estimated amplitude spectra for $n = 30$, 40, 70, and 100 from random walks of length 100,000. In figure 6 we show the results for $n = 100$.

The most striking feature of the amplitude spectrum of RNA landscapes is a strong difference between even and odd modes. This can easily be explained in terms of the physics underlying RNA folding: The major contribution of the folding energy comes from stacking of base pairs. Hence the major changes in free energy caused by a point mutation will arise from these contributions. Since stacking energies are influenced by even number of nucleotides depending on the location of the affected base pair within a stack. A recent comparison of amplitude spectra for different landscapes based on folding very short RNA chains indicates that the amplitude spectra of the free energy landscapes are typical (Stadler, 1998).

Most of the early work on RNA free energy landscapes (Fontana *et al.*, 1991; Fontana *et al.*, 1993; Schuster and Stadler, 1994) uses the correlation function $\rho(d)$ in terms of the Hamming distance and defines the correlation length ℓ^* as the distance in which ρ drops to $1/e$. Since random walks occasionally take steps with decreasing Hamming distance from the starting point, we have $\rho(\tau) \leq r(\tau)$, as long as $\tau \leq n/2$, and the correlation lengths estimated from ρ are significantly smaller than the numbers obtained from the random walk correlation function.

Table 4. Estimated parameters for large **GC** free energy landscapes. The data for ℓ_ρ are taken from table 3 by Fontana *et al.* (1993). The value $\ell_\rho = 1.8$ for $n = 18$ is estimated from $\ell_\rho = 1.96$ for $n = 20$. ℓ^* is obtained from a quadradic approximation of $\hat{r}(s)$ in the range $0.5 > r(s) > 0.2$ and solving for $r(s) = 1/e$. This procedure underestimates the correlation length ℓ, as defined in equ.(6.1), by almost 50%. This is due to the fact that the low-p modes dominate the correlation functions while the contributions of modes with large p are already close to 0 when $r(s)$ crosses $1/e$. Note that the average mode, $\overline{p}$, is independent of the chain length within the numerical accuracy.

n	ℓ_ρ	ℓ^*	ℓ	$\overline{p}$
18	(1.8)	2.19	3.18	2.83
30	2.96	3.24	4.47	3.35
40	4.00	4.19	5.75	3.48
70	6.62	7.48	11.00	3.18
100	8.46	10.19	15.42	3.24

From the definition of the correlation length, equ.(6.1), we find for an elementary landscape $r(s) = (1-1/\ell)^s$ and hence $r(\ell^*) = (1-1/\ell)^{\ell^*} = 1/e$. This equation is easily solved for ℓ^*:

$$\ell^* = -\frac{1}{\ln(1-1/\ell)} = \ell - \frac{1}{2} + \mathcal{O}(1/\ell)\,. \tag{6.2}$$

Thus the correlation measures ℓ and ℓ^* give very similar results for elementary landscapes. This is no longer true when the amplitude spectrum is spread out. In fact, we find a discrepancy of almost 50% between ℓ and ℓ^* in the RNA free energy landscapes, see table 4.

It was noted by Fontana *et al.* (1993) that the correlation length must increase linearly with chain length: the free energy difference introduced by a single point mutation is bounded above by a constant (some 10kcal/mol) while the expected free energy increases linearly with n. As a consequence the nearest neighbor correlation $r(1) = 1 - \mathcal{O}(1/n)$ and hence $\ell = \mathcal{O}(n)$. Consequently, the average mode $\overline{p}$ must approach a constant value for large n. The numerical data compiled in table 4 shows that $\overline{p}$ is constant within the (considerable) numerical error for $n \geq 30$ and takes a value around 3.3

7. Discussion

The *amplitude spectrum* of a fitness landscape can be used to analyze complex problems in evolutionary theory and evolutionary computation. The basis of the approach taken here is the notion of a landscape defined over a configuration space. The topological features of the configuration space are determined by the (genetic) operators which transform types into each other. Since each operator can in principle induce a different topology, the resulting fitness landscape can be different for each operator. This is reflected in the *One Operator-One Landscape* hypothesis (Jones, 1995).

Taken literally, this thesis could imply that recombination acts on a completely different landscape than mutation. The consequence could be that mutation and recombination are pursuing different optimization paths, which would be deleterious if both are acting together, as is the case in most higher organisms. Nevertheless, the algebraic approach to landscape analysis reveals deep commonalities between string recombination and point mutation landscapes.

For instance, already the hypergraph representation of string recombination spaces has revealed that the mutation graph is embedded in the recombination space (Gitchoff and Wagner, 1996). Hence, the point mutation space is homomorphic to the string recombination space. In addition, the extension of Fourier decomposition to landscapes on recombination P-structures shows even deeper relationships between mutation and recombination spaces: Walsh functions form a Fourier basis in both cases.

In this contribution we have explicitly computed the amplitude spectra of three very different types of landscapes for both 1-point crossover and mutation. We have shown that, despite some numerical difficulties, amplitude spectra can be estimated reliably from random walk data.

A comparison of the results shows substantial differences among different types of landscapes. Consider the mutation case first. While the synchronizing-CA and the NK landscapes exhibit a more or less monotonically decreasing amplitude spectrum, RNA landscapes have a distint even/odd pattern. The latter can be explained easily in terms of the biophysical properties of RNA folding.

The more rugged CA landscapes have substantial contributions B_p for large p, while for NK models with $k = 2$ only the first 3 modes contribute. We also note that the NK and in particular the CA landscapes have substantial additive constributions B_1 which are lacking in the RNA landscapes. The additive part of the fitness function introduces a long range bias in the landscape distinguishing "lowlands" for "highlands".

The effect of the linear mode is much more prominent in the crossover case. In fact, all modes except for the additive one contribute approximately evenly to the landscape as seen by the 1-point crossover operators. The large value of B_1 for CA and NK in the recombination case indicates that a GA has "something to work with" and should be capable of producing good solutions. Not so in the RNA case. This landscape looks extremely rugged for recombination, and the B_1 mode is very small. This observation is consistent with a recent study on RNA-evolution under the influence of recombination which shows that crossover is in fact detrimental on many RNA landscapes (Ukrainczyk & Fontana, personal communication 1997).

Our results indicate that (the ruggedness of) fitness landscapes of practical interest are not adequately described by a single parameter such as the correlation length. The amplitude spectra considered in this contribution provide a good summary of a landscape's properties. From an informational point of view, the amplitude spectrum is equivalent to the knowledge of the entire correlation function $r(s)$. It allows for a much more detailed interpretation, however, because B_p directly measures the influence of the interaction order p on the fitness function. Different types of landscapes exhibit distinctive patters in the amplitude spectrum that can (at least in some cases) be explained directly in terms of the processes generating the landscape. We have seen that simplified approximations of the amplitude spectrum using only a few "characteristic modes" in earlier studies of both the synchronizing-CA and the RNA landscapes do not yield satisfactory results. Finally, simpler measures of ruggedness such as the correlation length or the effective interaction order can easily be computed from the amplitude spectrum.

Acknowledgments

This work was supported in part by NSF grant IRI-9320200 (WH).

References

Altenberg, L. and Feldman, M. W. Selection, generalized transmission, and the evolution of modifier genes. I. The reduction principle. *Genetics* **117**, 559–572 (1987).

Angel, E. and Zissimopoulos, V. *Towards a classification of combinatorial optimization problems relative to their difficulty for generalized local search.* Tech. Rep. 1112, LRI, Université de Paris-Sud (1997).

Angel, E. and Zissimopoulos, V. Autocorrelation coefficient for the graph bipartitioning problem. *Theor. Computer Sci.* **19**, 229–243 (1998a).

Angel, E. and Zissimopoulos, V. *On the landscape ruggedness of the Quadratic Assignment Problem.* Tech. Rep. 98-02, LIPN, Institute Galilée, Université de Paris-Nord (1998b).

Berge, C. *Hypergraphs.* Amsterdam NL: Elsevier (1989).

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

Bonhoeffer, S., McCaskill, J. S., Stadler, P. F. and Schuster, P. RNA multi-structure landscapes. a study based on temperature dependent partition functions. *Eur. Biophys. J.* **22**, 13–24 (1993).

Burden, R. L. and Faires, J. D. *Numerical Analysis*. Boston, MA: PWS-KENT Publishing Company, 4th edn. (1989).

Cairns, T. W. On the Fast Fourier Transform on finite abelian groups. *IEEE Trans. Computers* **20**, 569–571 (1971).

Cech, T. R. Conserved sequences and structures of group I introns: building an active site for RNA catalysis — a review. *Gene* **73**, 259–271 (1988).

Chung, F. R. K. *Spectral Graph Theory*, vol. 92 of *CBMS*. American Mathematical Society (1997).

Codenotti, B. and Margara, L. *Local Properties of Some NP-Complete Problems*. Tech. Rep. TR-92-021, International Computer Science Institute, Berkley, CA (1992).

Culberson, J. C. Mutation-crossover isomorphism and the construction of discriminating functions. *Evol. Comp.* **2**, 279–311 (1995).

Das, R., Crutchfield, J. P., Mitchell, M. and Hanson, J. E. Evolving globally synchronized cellular automata. In: *Proceedings of the Sixth International Conference on Genetic Algorithms* (Eshelman, L. J., ed.), pp. 336–343. Morgan Kaufmann (1995).

Eigen, M., McCaskill, J. and Schuster, P. The molecular Quasispecies. *Adv. Chem. Phys.* **75**, 149–263 (1989).

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

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

García-Pelayo, R. and Stadler, P. F. Correlation length, isotropy, and meta-stable states. *Physica D* **107**, 240–254 (1997).

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

Gitchoff, P. and Wagner, G. P. Recombination induced hypergraphs: A new approach to mutation-recombination isomorphism. *Complexity* **2**, 47–43 (1996).

Goldberg, D. E. Genetic algorithms and walsh functions. Part I: a gentle introduction. *Complex Systems* **3**, 129–152 (1989).

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

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

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

Happel, R. and Stadler, P. F. Canonical approximation of fitness landscapes. *Complexity* **2**, 53–58 (1996).

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

Hordijk, W. A measure of landscapes. *Evol. Comp.* **4**, 335–360 (1996).

Hordijk, W. Correlation analysis of the synchronizing-ca landscape. *Physica D* **107**, 255–264 (1997).

Jones, T. *Evolutionary Algorithms, Fitness Landscapes, and Search*. Ph.D. thesis, Univ. of New Mexico, Albuquerque, NM (1995).

Kauffman, S. A. Adaptation on rugged fitness landscapes. In: *Lectures in the Sciences of Complexity* (Stein, D., ed.), pp. 527–618. Addison-Wesley (1989).

Kauffman, S. A. *The Origin of Order*. New York, Oxford: Oxford University Press (1993).

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

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

Lyubich, Y. I. *Mathematical Structures in Population Genetics*, vol. 22 of *Biomathematics*. Berlin: Springer-Verlag (1992).

Maslen, D. and Rockmore, D. Generalized FFTs – a survey of some recent results. In: *Groups and Computation II* (Finkelstein, L. and Kantor, W., eds.), vol. 28 of *DIMACS*, pp. 183–238. Providence, RI: American Mathmatical Society (1996).

Mézard, M., Parisi, G. and Virasoro, M. *Spin Glass Theory and Beyond*. Singapore: World Scientific (1987).

Mohar, B. The Laplacian spectrum of graphs. In: *Graph Theory, Combinatorics, and Applications* (Alavi, Y., Chartrand, G., Ollermann, O. and Schwenk, A., eds.), pp. 871–898. New York: John Wiley and Sons, Inc. (1991).

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

Rockmore, D. Some applications of generalized FFTs. In: *Groups and Computation II* (Finkelstein, L. and Kantor, W., eds.), vol. 28 of *DIMACS*, pp. 329–370. Providence, RI: American Mathmatical Society (1995).

Schuster, P. and Stadler, P. F. Landscapes: Complex optimization problems and biopolymer structures. *Computers & Chem.* **18**, 295–314 (1994).

Schuster, P., Stadler, P. F. and Renner, A. RNA structures and folding: From conventional to new issues in structure predictions. *Curr. Opinions Structural Biol.* **7**, 229–235 (1997).

Sorkin, G. B. *Combinatorial optimization, simulated annealing, and fractals*. Tech. Rep. RC13674 (No.61253), IBM Research Report (1988).

Stadler, P. F. Correlation in landscapes of combinatorial optimization problems. *Europhys. Lett.* **20**, 479–482 (1992).

Stadler, P. F. Towards a theory of landscapes. In: *Complex Systems and Binary Networks* (Lopéz-Peña, R., Capovilla, R., García-Pelayo, R., Waelbroeck, H. and Zertuche, F., eds.), pp. 77–163. Berlin, New York: Springer Verlag (1995).

Stadler, P. F. Landscapes and their correlation functions. *J. Math. Chem.* **20**, 1–45 (1996).

Stadler, P. F. Fitness landscapes arising from the sequence-structure maps of biopolymers. *J. Mol. Struct. (THEOCHEM)* (1998). In press, Sante Fe Institute Preprint 97-11-082.

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

Stadler, P. F. and Happel, R. Random field models for fitness landscapes. *J. Math. Biol.* (1998). In press, SFI preprint 95-07-069.

Stadler, P. F. and Schnabl, W. The landscape of the traveling salesman problem. *Phys. Letters A* **161**, 337–344 (1992).

Stadler, P. F. and Wagner, G. P. The algebraic theory of recombination spaces. *Evol. Comp.* **5**, 241–275 (1998).

Tacker, M., Stadler, P. F., Bornberg-Bauer, E. G., Hofacker, I. L. and Schuster, P. Algorithm independent properties of RNA structure prediction. *Eur. Biophy. J.* **25**, 115–130 (1996).

Wagner, G. P. and Stadler, P. F. Complex adaptations and the structure of recombination spaces. In: *Algebraic Engineering* (Nehaniv, C. and Ito, M., eds.). Singapore: World Scientific (1998). Proceedings of the Conference on Semi-Groups and Algebraic Engineering, University of Aizu, Japan, in press. Santa Fe Institute Preprint 97-03-029.

Weinberger, E. D. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biol. Cybern.* **63**, 325–336 (1990).

Weinberger, E. D. Fourier and Taylor series on fitness landscapes. *Biol. Cybern.* **65**, 321–330 (1991).

Welch, L. A computation of finite Fourier series. *JPL Space Programs Summary* **4** (31–37), 295–296 (1968).

Wright, S. The roles of mutation, inbreeding, crossbreeeding and selection in evolution. In: *International Proceedings of the Sixth International Congress on Genetics* (Jones, D. F., ed.), vol. 1, pp. 356–366 (1932).

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